

MEASUREMENT  
AND EVALUATION  
IN PSYCHOLOGY  
AND EDUCATION

Copyright, 1955 © 1961 by John Wiley & Sons, Inc.

*All rights reserved. This book or any part thereof,  
must not be reproduced in any form without  
the written permission of the publisher.*

Library of Congress Catalog Card Number: 61-11494

Printed in the United States of America

# Preface

The reception that this book received in its first edition has been sufficiently favorable, and the comments that have come back to us have been sufficiently kind, that we have tried to make this revision an evolution rather than a revolution.

We have, of course, tried to make adequate reference to major new tests that have appeared in the last six years—*Form L-M* of the *Stanford-Binet*, *WAIS*, *STEP*, and others. We have also tried to make some reference to significant recent research, where it fits the pattern of an introductory text.

Beyond that, we have redone a number of sections from the first edition—sections with which we or our users had been less than completely happy. Thus, the section on validity has been rather completely restructured to represent our current thinking. A fuller exposition has been developed on planning and blueprinting a test. We have dealt more fully with the practical mechanics of a testing program. We have reorganized and added to the material on aptitude-test batteries.

In two instances, we have changed the order of chapters into a sequence that seems to us more teachable. However, whenever information is presented in sequence, one feels the need for what is to come later while one is discussing what has come before. Some who have used the book in classes have told us that they teach the chapters in various sequences different from that which appears in the book, and, fortunately, this seems to lead to no insuperable problems.

We continue in the basic belief that it is as important for students to learn what tests will *not* do as to learn what they *will* do; as important to examine their own purposes and objectives for testing as to examine the tests. It is in the hope of developing more restrained, discriminating, and insightful testers that we offer this book to our colleagues and students.

ROBERT L. THORNDIKE  
ELIZABETH HAGEN

New York, N. Y.  
April, 1961

# Contents

CHAPTER	
1. Historical and Philosophical Orientation	1
2. Overview of Measurement Methods	17
3. The Teacher's Own Tests	27
4. Preparing Objective Tests	60
5. Elementary Statistical Concepts	96
6. Norms and Units for Measurement	124
7. Qualities Desired in Any Measurement Procedure	160
8. Where to Find Information about Specific Tests	207
9. Standardized Tests of Intelligence or Scholastic Aptitude	219
10. The Measurement of Special Aptitudes	261
11. Achievement Tests	288
12. Questionnaires and Inventories for Self-Appraisal	317
13. The Individual as Others See Him	351
14. Behavioral Measures of Personality	388
15. Projective Tests	422
16. Planning a School Testing Program	444



CHAPTER	
17. Marking and Reporting	484
18. Measurement in Educational and Vocational Guidance	521
19. Tests in the Selection of Personnel	542
APPENDIX	
I. Computation of Square Root	565
II. Calculating the Correlation Coefficient	567
III. <i>Section A</i> General Intelligence Tests	571
<i>Section B</i> Aptitude Test Batteries	575
<i>Section C</i> Reading Tests	577
<i>Section D</i> Elementary-School Achievement Batteries	582
<i>Section E</i> High-School Achievement Batteries	584
<i>Section F</i> Interest Inventories	586
<i>Section G</i> Adjustment and Temperament Inventories	588
IV. Sources for Educational and Psychological Tests	592
Index	597

## Chapter 1



# Historical and Philosophical Orientation

### HISTORICAL BACKGROUND

The roots of the measurement of man lie in antiquity. We must believe that even in prehistoric times Og, the cave man, made rudimentary appraisals of his fellows. He saw Zog go by, made some such judgment as "Big, strong, keep out of way," and acted upon it; or he came upon the campfire of Wog, observed "Small, weak, take dinner," and did so forthwith. But for much of recorded history, the appraisals that man has made of his fellows have been of this crude subjective type.

He who seeks imaginatively can find suggestions of more systematic and refined methods. Thus, the tournaments of the days of chivalry can be thought of as an effort to arrange men in an order from best to worst in feats of arms, and contests leading to the crowning of "champions" have always constituted a rough sort of measurement. Teachers have always catechized their pupils to appraise their degree of mastery of the tasks assigned them, evaluating them as best they could by their responses. But these approaches were more primitive than the sun dial and the ox cart. They are characteristic of the appraisal of man and his behavior up to the present century. Application of the quantitative methods of science to psychology and education is very new. In 1850 there was almost none of it; 1900 was still a pioneering period.

### EARLY EDUCATIONAL TESTING

The appraisal of educational achievement in the United States before 1850 had relied very largely upon oral examination. The teacher or visiting examiner asked a question. The designated pupil undertook to answer it. The questioner arrived at an immediate subjective evaluation of the answer. There was uniformity neither in the ques-

tions asked different pupils nor in the evaluation of their replies. The method was burdensome and inefficient, since only one pupil could be tested at a time. It provided no comparability from pupil to pupil either in the task or in the evaluation of it.

During the latter half of the nineteenth century, oral examinations by boards of visitors were replaced by set written examinations as a basis for promotion or admission to an academy or college. Outside examination in turn yielded to evaluation by the classroom teacher. Whether carried out by an outside examiner or by a teacher, however, the technique was that of the essay examination, in which a pupil responded in his own words to a question set by the examiner.

The written examination had advantages over the oral examination of (1) presenting the same tasks to each member of the group and (2) letting each pupil work for the full examination period. However, though the task was made uniform, at least for the members of a given class, appraisal of each individual's response to the task remained highly subjective, depending upon the standards and prejudices of the particular scorer. As we shall see in Chapter 3, great variations were found in the *scoring* of a particular paper. Only since 1900 has there been any general development of objectively scored tests in which a pre-established key can be routinely and uniformly applied to the responses made by each pupil. Only since 1900 has the idea emerged of a general standard of performance for an age or grade, with which the performance by any class or any individual may be compared.

#### THE BEGINNINGS OF PSYCHOLOGICAL MEASUREMENT

Psychology in 1850 was still in large measure a part of philosophy. Courses dealing with man and his actions were presented under the title "Moral Philosophy," and discussed in an armchair fashion the nature of the Mind and the Soul. Psychology was almost entirely non-experimental, and the idea that one could measure in quantitative terms the speed of responding, the amount of forgetting, or the level of intelligence would have been received in most quarters with hostility or, more probably, ignored as not worthy of rebuttal. The nearest approaches to psychological measurement were a few scattered experiments by physicists and physiologists on the measurement of the ability to make sensory discriminations and the speed of simple elementary responses.

By 1900 psychology had felt the impact of the physical and biological sciences and was striving mightily to become a science itself. It was shaking off the ties that bound it to philosophy and forming new

alliances with the biological sciences. It had adopted the experimental method and was measurement-conscious. The basic tool of experimentation is measurement, and psychology was expanding its measurement techniques in all directions. The record since 1900 is the record of the attempt to expand and adapt measurement techniques to cover all aspects of human behavior.

Three main streams combined to yield the vigorous measurement movement in psychology and its spread through education. Some of the flavor and some of the emphasis have come from each stream. These were (1) the physiological and experimental psychology that had its main growth in Germany in the nineteenth century, (2) Darwinian biology, and (3) the clinical concern for the maladjusted and underdeveloped individual.

#### BEGINNINGS OF EXPERIMENTAL PSYCHOLOGY

The modern scientific era was first ushered into the physical sciences in the seventeenth and eighteenth centuries. Scientific interest and method soon spread over to the biological sciences, and by the early nineteenth century experimental physiology was a center of active research interest in the experimental laboratories in Germany and other European countries. Experimental physiologists became interested in the operation of the senses, studying intensively seeing, hearing, and the other senses. Physiologists also became interested in measuring the speed of simple motor responses.

In 1879 the first laboratory for experimental psychology was established by Wilhelm Wundt at Leipzig. Early experimental psychologists were interested in many of the same measurements that had concerned the physiologists. These were measures of seeing, hearing, feeling, and speed of response. But gradually they extended their concern to more clearly psychological matters, such as measurement of perceptual span—the amount that the individual can “take in” at once, of rate of learning, of the timing of complex mental tasks, and so forth.

One area of particular interest for its contribution to the broad field of psychological and educational measurement was that known as *psychophysics*. The experimental psychologist was much interested in exploring the relationship between physical stimulus intensities, e.g., of light wave or of sound wave, and the experienced intensity of the resulting sensation. The designing of effective experimental procedures for studying these problems gave rise to a set of techniques that have proved adaptable to a wide range of problems of psychological measurement.

From experimental psychology came a legacy of respect for careful experimental method and precision of technique, a number of experimental designs, and statistical techniques that could be carried over to more general psychological and educational measurement problems.

#### EARLY STUDY OF INDIVIDUAL DIFFERENCES

A second stream contributing to psychological measurement was Darwinian biology. In 1859 Darwin brought out his *Origin of Species*. The basic concern in Darwin's work was with variation among the members of a species, that is, individual differences. Darwin's work was followed up in England and applied to distinctively human affairs, particularly by Sir Francis Galton. Whereas German psychology had focused on finding the general facts true of all people, Galton became interested primarily in the differences among people. Stimulated by Darwin to study the inheritance of traits, he gathered data both on physical and on psychological characteristics. The study of these individual differences required better statistical tools, and the British group, under the leadership of Karl Pearson, developed improved techniques for analyzing and describing the patterns of individual differences.

These, then, were the two main contributions of the British group to the growth of psychological measurement: a deep concern for studying the differences among people as interesting and significant facts and appropriate statistical techniques and tools for carrying out this study.

#### CLINICAL STUDY OF DEVIATES

During this same period, a third stream was gathering strength. This was concern for the individual who was not functioning successfully. Humanitarian concern for the insane, the feeble-minded, and the general misfit led in the nineteenth century to active research and investigation aimed toward understanding their condition and improving their lot. This clinical interest in the maladjusted individual was particularly strong in France, and it was here that it bore fruit for the field of measurement. As psychologists worked with these unfortunate deviates, the need became more and more apparent for some uniform way of expressing the degree of their defect, particularly in the mental sphere. It was in this context of concern for the child who was not getting along in school that Binet and his colleagues developed the series of intellectual tasks that ultimately grew into the whole array of measures of intelligence.

## SYNTHESIS IN THE UNITED STATES

By the early years of the present century, all these streams of influence had made themselves felt in the United States. James McKeen Cattell had taken his graduate work in psychology in Germany with Wundt, where he had received a good grounding in quantitative and experimental psychology. But he had also been exposed to the work of Galton and had developed a lasting interest in individual differences and statistical method. When he returned to the United States, he began an investigation of individual differences in the simple sensory and motor performances that were being measured in German psychological laboratories. He studied the relationship between these performances and academic success.

E. L. Thorndike was a student of Cattell's just before the turn of the century and became a focal influence in the spread and development of standardized educational tests. Both his own work and that of a large group of students rapidly spread the gospel of objective measurement in education.

The work of Binet was eagerly seized upon in this country. His tests were translated and produced in several versions, of which by far the most influential became the *Stanford-Binet* first produced by Lewis Terman in 1916. The testing movement seemed especially suited to the temper of this country and took hold here with a vigor and enthusiasm unequaled elsewhere.

## MEASUREMENT IN THE TWENTIETH CENTURY

The first 60 years of the twentieth century may conveniently be divided into four equal parts, so far as the recent history of psychological and educational measurement is concerned. We may designate the period from 1900 to 1915 the pioneering phase. This was the period of exploration and initial development of methods. It saw the emergence of the first Binet intelligence scales and their American revisions. Standardized achievement tests in different subjects began to appear, exemplified by Stone's arithmetic tests, Buckingham's spelling tests, and Trabue's language tests. Thorndike developed his first handwriting scale. Otis and others were initiating work on group tests of intelligence.

The next 15 years, 1915 to 1930, can perhaps be called the "boom" period in test development. The pioneers had shown the way, and in the hands of enthusiastic followers tests multiplied like rabbits. Standardized tests were developed for all the school skills and for the content areas of the school program. Achievement batteries made their appearance. Starting with *Army Alpha* of World War I, group

intelligence tests were produced in great numbers. Also starting with a wartime product, the *Woodworth Personal Data Sheet*, a whole line of personality questionnaires and inventories came into being.

The rapid development of testing instruments and methods was pushed by a group of enthusiasts. They were converts who had "gotten the word." Their enthusiasm was contagious and extended not only to the production of tests but also to their use. Tests of intelligence and achievement were administered widely and somewhat indiscriminately. Test results were often accepted unhesitatingly and uncritically and served as the basis for a variety of unjustified judgments and actions with respect to individuals. In the expansive flood of enthusiasm for objective measurement, some enthusiasts were not inclined to be critical of their instruments or the interpretation of results from them. Many sins were committed in the name of measurement by uncritical test users.

After a while the pendulum began to swing back. More and more sharply voiced criticisms of tests and of the uses made of tests began to be heard. Heredity-environment discussions became acrimonious. The use of test scores as a basis for classroom grouping became the subject of bitter attack. Criticism was directed at specific tests in terms of their limited scope and their emphasis upon restricted and traditional objectives. It was also directed at the whole underlying philosophy of quantification and the use of numbers to express psychological qualities.

The critical attack had the healthy effect of forcing the test enthusiasts themselves to become more critical of their assumptions and procedures and to broaden their approach to the whole problem of psychological and educational appraisal. From about 1930 to 1945 may be considered a period of critical appraisal, of taking stock, of broadening techniques and delimiting interpretations. It was a period in which the center of attention shifted from "measuring" a limited range of academic skills to "evaluating" achievement of the whole range of educational objectives. It was a period in which the holistic, global projective methods of personality appraisal came to the fore.

It is difficult to view with any perspective at all events that have taken place within the last 15 years. History may eventually characterize the period quite differently than do we, standing so close to it. However, we will venture to predict that the period from 1945 to 1960 will be characterized as the period of test batteries and testing programs. Partly as a result of their successful use in World War II, integrated aptitude batteries for educational and personnel use have multiplied during this period. And the large-scale testing programs,

such as those administered by the College Entrance Examination Board, though stemming from much earlier in the century, have expanded in size and multiplied in numbers at a striking rate. We have experienced a *second boom period*—not so much in test development and construction, as in test administration and use. The mid-twentieth century is a period in which standardized testing is a widely experienced and widely accepted phenomenon of our American culture.

Under these circumstances it is particularly important that construction, use, and interpretation of these instruments be well understood by teachers, guidance workers, and psychologists for whom they are daily tools of the trade. It is also important that the phenomenon of standardized testing be understood by the citizens who are exposed to it in their search for employment for themselves or education for their children. Therefore, let us try at this point to formulate a philosophy of measurement that will take into account the lessons of the last 60 years, and will serve to guide our attack on measurement problems and our use of measurement techniques in the years ahead.

## PHILOSOPHICAL ORIENTATION

In education and in psychology we are concerned with human beings. Sometimes we are concerned with them as specific individuals, as when we want to know why Mary is having so much difficulty in learning to do long division. Sometimes we are concerned with them as specific groups of individuals, as when we inquire whether the children in class A can read as well as those in class B. Sometimes we are concerned with them as general representatives of mankind, as when we try to determine whether children with high verbal intelligence tend to show more or less signs of emotional disturbance than children of average intellectual ability.

### KNOWLEDGE AS A GUIDE TO ACTION

In practically all of education and in much of psychology, our concern about individuals is to *do* something about them, individually or collectively. In so far as education is a science, it is an applied science, and in psychology, too, the applied aspects bulk large in the present scene. The educator or the practical psychologist is continually faced with the necessity of arriving at some decision as to a course of action. He must decide what to do about an individual or individuals, or he must help the person himself decide what to do. He must decide in which grade to place a child or what special in-



struction to provide for him. He must reach a diagnosis of a child with a reading disability, with a view to recommending treatment. He must recommend whether or not to employ a job applicant. He must help a student decide whether to plan for college and, if so, what sort of program to take and what type of job to aim for. The educator or psychologist wants each one of these decisions to be a sound and well-conceived one.

Our basic assumption is that *sound decisions arise out of relevant knowledge* of the individual or individuals. We assume that the more we know about a person that relates to our present decision, and the more accurately we know it, the more likely we are to arrive at a sound decision about him or a wise plan of action for him. By the same token, we assume that the more relevant and accurate information we can provide the individual about himself, the more likely he is to arrive at a sound decision on his own problem. It may be necessary for us to qualify this assumption as we proceed. There may be limits on the amount and kind of information that can be used in a particular situation. We shall indicate that knowledge in and of itself is not wisdom. But in its general form the assumption is basic not only to educational and psychological measurement but also to all science. We assume basically that knowledge is good, that knowledge is power, that knowledge is the basis for effective control of the problems that confront us from day to day. This is a basic tenet of our faith.

What does it mean to "know an individual"? Fundamentally, to know an individual means to be able to describe him accurately and fully. If we know John Jones well, we can describe not only how he looks—how tall he is and how heavy, the color of his hair and eyes, the birthmark under his chin. Much more importantly, we can describe what he can and will do—how he will dress, what he is likely to talk about, what he will be interested in, what types of tasks he can do and how well he can do them, how he will respond to the different stresses and strains of life. To know a person completely means to be able to describe him completely, to predict how he will behave in every possible situation. Obviously, we are far, far away from this objective, and we always will be. The function of educational and psychological measurement is to move us a little closer to it.

#### IMPORTANCE OF MEASURING THE RIGHT THING

The effectiveness of our description of any object or person depends upon two things. It depends (1) upon how wisely we have chosen

the features to be described and (2) upon how truly and accurately we have managed to describe each one.

A description may fail to be useful for the need at hand because we choose irrelevant features to describe. Thus, in describing a painting we might report its height, its breadth, and its weight. We might report these with great precision. If our concern were to crate the picture for shipment, these might be just the items of information we would need. On the other hand, if our purpose was that of characterizing the painting as a work of art, our description would be worthless. The attributes of the picture we had described would be essentially irrelevant to its quality as a work of art.

Similarly, a description of a person may be of little value for our purposes if we choose the wrong things to describe. Thus, the Air Force in selecting pilots to fly jet fighters might get very accurate information on height, weight, years of education, size of vocabulary, and speed of reading for all its applicants. It would almost surely find, however, that none of these things helped at all in selecting the men who could successfully learn to fly the planes. Such factors as these are in large measure irrelevant to flying success, which appears to depend more on mechanical know-how and on motor coordination.

Again, a high school concerned about assessing the level of literary appreciation in its pupils might prepare a test inquiring exhaustively into the names of the characters and the details of the plot of Shakespeare's *Julius Caesar*. The worthlessness of this procedure may be less obvious, but is probably just as real as that proposed for the selection of pilots. This test seems useless for the task at hand, because detailed factual knowledge about an isolated literary work is no indicator of the quality of a pupil's literary appreciation. The test has asked the wrong kinds of questions. The evidence it provides is related to a faulty interpretation of the original question that was asked.

The first, and perhaps the most important, step in any project for educational or psychological measurement is defining just what it is that we wish to measure and determining what operations will serve to measure it. Educational objectives are likely to be incompletely formulated and expressed in vague terms. The concepts must be clarified and made more specific before we can make much progress toward sensible procedures of measurement. Until we can decide what is meant by "good citizenship" or what behaviors are exhibited by a person who shows "understanding of scientific method," we have little prospect of developing procedures to appraise either the one or the other.

## THE NEED FOR PRECISION

Our description may be of limited value in the second place because the attributes we elect to describe are described inaccurately. Thus, if our description of the painting were expressed in terms of theme, composition, line and volume, and color values, it would certainly be a good deal more to the point as an appraisal of a work of art. But it would be much more wordy, more subjective, and less precise than our previous description of length, breadth, and weight. Different persons could be expected to differ markedly in the qualities they saw and the terms they used to describe them. This might be true to such an extent that a single individual's description would give us only a very rough, unclear, and undependable impression of the picture as a work of art.

As for the candidate for pilot training, we might get ratings from his friends on his speed of learning new coordinations, ability to pay attention to many things at once, and resistance to disturbance by emotional stress. We may hazard a guess that these ratings would again prove ineffective in predicting pilot success—not so much because the qualities themselves are unimportant, but because we are not skillful in observing such qualities in our fellows or in expressing our observations in exact quantitative form.

Our high school, concerned with literary appreciation, might ask each pupil to write a report on some book he had read recently, telling what he had liked about it, and why he had or had not thought it was a good book. Again, we may feel that such a report would provide information more related to appreciation than would a test of factual knowledge. But judging the quality of appreciation shown in a varied collection of compositions about an assortment of different books would be a very subjective enterprise, and the judgments would tend to be quite undependable. Each judge would have his own personal standards of what constituted good literary appreciation. He would make his judgments in terms of those personal standards. There would be little agreement from one judge to the next as to who had shown good appreciation and who had shown poor. Our appraisal would be unsatisfactory because it would be inaccurate.

## DEGREES OF REFINEMENT IN MEASURING

There is enormous variation from one trait to another in the degree of refinement we have been able to achieve in describing it. At the crudest level, our appraisal may come to no more than a simple two-way classification. This may take the form present—absent: e.g.,

John lisps but Bill does not lisp; or the form trait—opposite: e.g., John runs fast but Bill runs slowly.

A somewhat more refined level of description is achieved when we characterize the trait by a set of adjectives which represent degrees of the trait: e.g., John runs fast, Joe goes like a streak, Jack runs fairly fast, Will goes like molasses in January. But the number of such qualitative descriptions is limited, and the meaning of such adjectives or similes is far from uniform from person to person.

A still further level of refinement in description is reached when we can arrange the members of a group in rank order with respect to an attribute and when we can locate any individual on such a rank order. Thus, we may say Joe runs the fastest, John runs faster than Jack, Jack runs faster than Will, and Will runs the slowest. Such a procedure of ranking could theoretically be extended to include all the children in a class, in a school, or even all the children in the whole country. Clearly, when we can appraise a trait well enough to produce such a ranking, a very great increase in the adequacy of our description has been achieved.

Finally, some attributes may be expressed in a quantitative statement of amount. Thus, we may be able to report that Joe ran 100 yards in 10 seconds, John in 14 seconds, Jack in 15 seconds, and Will in 17 seconds. This last is clearly the most precise type of statement of the essential facts and the one that makes us best able to decide upon appropriate action with regard to an individual, so far as that action depends upon speed of running 100 yards. It is certainly the type that the track coach would want to have before deciding whom to keep on the track team.

We have identified four points along a scale of quantification and precision of measurement.

1. *Either—Or.* A pupil is either a boy or a girl. A man is single, married, widowed, or divorced. A student is enrolled in the college preparatory, commercial, or general curriculum.

2. *Qualitatively Described Degrees.* Thus, a pupil may show "normal speech," "slight stuttering," "stammer," "marked stutter." Or the pupils in a class may be characterized as "quiet and relaxed," "slightly fidgety," or "tense and restless."

3. *Rank in a Group.* Thus, a series of graded tasks scored by uniform standards enables us to find who does best and who does worst on reading comprehension, arithmetic problems, or spelling. The rest of the group can be arranged in order from best to worst.

4. *Amount, Expressed in Uniform Established Units.* A boy weighs 56 pounds, is 45 inches tall, is  $6\frac{1}{2}$  years old.

This wide variation in the refinement of our appraisals must be frankly admitted. Some traits we may never be able to express more accurately than by a "very" and "not very" characterization. Our failure to have achieved greater refinement in measuring these traits is probably partly due to lack of clarity and sharpness in definition of the attribute that we propose to describe. When we characterize a person as sincere, cultured, socially adjusted, cooperative, a good citizen, our hearer may have only a very general idea of what we mean. (And, as a matter of fact, so may we.) In part our failure is certainly due to the limited ingenuity and skill we have shown to date in finding ways to represent degree or amount of the attribute with precision. It may sometimes be partly due to the essential nature of a particular attribute, which makes it fundamentally not expressible in quantitative form. There may be some things that, in their very nature, can never be quantified.

Certainly, our present ability truly to *measure* many of the attributes of persons that appear to be relevant and important for making decisions about them and planning actions with respect to them leaves much to be desired. However, while recognizing this fact we must also appreciate that enormous strides have been made since 1900 toward more objective and more accurate appraisals of human beings. The fact that we are limited in some directions does not lessen the value of increased precision wherever such increased precision has been achieved. While keeping a critical eye upon the limitations of measurement procedures, we should still use them for all they are worth in increasing the accuracy of our information about students, employees, or clients.

#### CRITICISMS OF PSYCHOLOGICAL AND EDUCATIONAL MEASUREMENT

Since about 1930, psychological and, particularly, educational measurement have come in for a good deal of criticism. The educational philosophers have been especially outspoken in expressing their dissatisfaction. In part, the criticisms have been directed at the basic logic of psychological measurement. These criticisms have been directed at the limitations we have just been discussing, as well as at some other problems concerning the equivalence of units and scores, which we shall consider briefly in a later chapter. In part, however, the criticisms have been directed at the effects that the measurement

procedures have had upon school practice. The following types of criticisms have been made:

1. *Standardized measurement procedures* have been said to foster undemocratic practices and attitudes in the classroom. Forming homogeneous class groups on the basis of an intelligence or achievement test is a specific practice that has been the target of this criticism.

2. It has been contended that standardized tests have had the effect of freezing the curriculum and of preventing experiment and change, on the grounds that the commercial standardized test typically lagged behind the advance of educational thought and practice.

3. The limited scope of many standardized tests has been pointed out, and it has been indicated that they fail to appraise many of the changes in children that schools should be interested in producing.

4. The short-answer test items have been accused of producing undesirable study habits directed toward piecemeal memorization rather than understanding.

There has been at least a germ of truth in all these criticisms. Some of them we shall consider in more detail in later chapters. As the criticisms are examined, however, we find that they are not primarily criticisms of obtaining more information about the individual. They are criticisms either of (a) the incompleteness and imperfection of the information yielded by our measurement procedures or (b) the unwise things that we do with that information. It is as though we condemned the physicist because (a) he cannot yet control the weather, and (b) his knowledge has led to the construction of atom bombs which may destroy mankind. We must grant that our measurement procedures are not complete and our actions based on them are not always wise. But the remedy lies in developing better measurement procedures that will give us more complete and more accurate information about the individual. It lies in gaining better understanding of our measures—their strengths and their weaknesses—so that we may use them with more wisdom. It does *not* lie in getting less information.

It cannot be too much emphasized that *measurement at best provides only information, not judgment*. A test will yield only a score, not the conclusion to be drawn from that score. The information provided in a test score is not a substitute for insight. This information is the raw material with which insight must work, in the clinic, in the classroom, and in the research laboratory. Experience, training, and basic sagacity must provide the insight that will take a set of data about an individual or group, know how much faith to place

in them and what meaning to give them, and draw from them a sound conclusion or plan for action.

Furthermore, it should be emphasized that the information that *any* measurement procedure gives is limited. It is limited by the nature of the measurement instrument itself. The typical intelligence test, for example, samples certain types of performances with abstract ideas expressed in symbolic form. It is not a measure of the general worth of the individual, of his ability to acquire mechanical skills or artistic techniques, or of his integrity and dependability as a member of society. The information is limited by the conditions under which the procedure is applied. Thus, an intelligence test given to an emotionally disturbed and resistant child may give a very inadequate picture of what that same child could do if the disturbing influences were removed and the resistance overcome. Learning to use measurement results wisely is in part learning what information a particular device does and does not provide and in part learning under what circumstances that information is likely to be trustworthy. Throughout this book there will be recurring attempts to guide that learning.

### SUMMARY STATEMENT

We can summarize much of the foregoing discussion on a working philosophy of measurement in the following four points.

1. The process of measurement is secondary to that of defining objectives. The ends to be achieved must first be formulated clearly. Then measurement procedures can be sought as tools for appraising the extent to which those ends have been achieved.

2. Much of educational and psychological measurement is, and will probably remain, at a relatively low level of precision. We must recognize this fact, using the best procedures available to us, but always treating the resulting score as a tentative hypothesis rather than as an established conclusion.

3. The more elegant procedures of formal test and measurement must be supplemented by the cruder procedures of informal observation, anecdotal description, and rating if we are to obtain a description of the individual that is usefully complete and comprehensive.

4. No amount of ingenuity in developing improved procedures for measuring and appraising the individual will ever eliminate the need to *interpret* the results from those procedures. Measurement procedures are only tools. Insight and skill are required in the use of such

tools. The sharper and more varied the tools, the more skill it takes to use them most effectively.

# SUGGESTED ADDITIONAL READING

- Boring, Edwin G., H. S. Langfeld, and H. P. Weld, *Foundations of psychology*, New York, Wiley, 1948, Chapter 18.
- Cottle, William C. and N. M. Downie, *Procedures and preparation for counseling*, Englewood Cliffs, N. J., Prentice-Hall, 1960, pp. 158-162; 165-167; 174-175; 180-183.
- Goodenough, Florence L., *Mental testing*, New York, Rinehart, 1949, pp. 3-96.
- Harris, Chester W., Editor, *Encyclopedia of educational research*, 3rd ed., New York, Macmillan, 1960, pp. 807-816; 1502-1503.
- Lorge, Irving, The fundamental nature of measurement, Chapter 14 in E. F. Lindquist, Editor, *Educational measurement*, Washington, D. C., American Council on Education, 1951.
- Murphy, G., *An historical introduction to modern psychology*, rev. ed., New York, Harcourt, Brace, 1949, Chapters 6, 8, 11, 24, and 26.
- Nunnally, Jum C., *Tests and measurements*, New York, McGraw-Hill, 1959, Chapters 1 and 2.
- Seward, Georgene S., and John P. Seward, *Current psychological issues*, New York, Holt, 1958, Chapter 11.
- Wrightstone, J. Wayne, et al., Educational measurements, *Rev. educ. Res.*, 26, 1956, 268-291.
- Wrightstone, J. Wayne, Joseph Justman, and Irving Robbins, *Evaluation in modern education*, New York, American Book, 1956, Chapter 1.

# QUESTIONS FOR DISCUSSION

1. The development of objective and standardized tests has proceeded faster and further in the United States than in any other country. What factors do you see as contributing to this?
2. Try to talk to a student from some foreign country and find out what examinations are like and how they are used in his country. What differences do you find, as compared with the United States? What are the advantages and disadvantages of each system?
3. In many graduate schools oral examinations are still used in examining candidates for higher degrees. What are the advantages and disadvantages of this type of examination?
4. From your reading or from your personal experience, give one or more concrete examples of the misuse or misinterpretation of the results from standardized tests.
5. How universally acceptable is the statement "knowledge is good" in the field of education and applied psychology? What objections would you have to this statement, or what limitations would you place upon it?



6. Give an illustration of a measuring procedure in education or psychology that would be of little or no value because it was not sufficiently precise; one that would be of no value because it was measuring the wrong thing.

7. Give two examples of educational or psychological measures to represent each of the following four points along the scale of quantification and precision of measurement: (a) *either—or*, (b) *qualitatively described degrees*, (c) *rank in a group*, (d) *amount, expressed in uniform, established units*.

8. Your textbook states that "to know an individual means to be able to describe him accurately and fully." What would be central in such a description for

- a. A fourth-grade girl having difficulty with arithmetic.
- b. An eighth-grade boy who has been picked up for throwing rocks through the school windows.
- c. A recent high-school graduate who is being considered for a job as receptionist.

## Chapter 2



# Overview of Measurement Methods

During the present century techniques for appraising the individual have been developed in great variety, and they have been applied to many aspects of his abilities and personality. Specific techniques will be discussed in detail in later chapters. The present chapter is devoted to a general overview, mapping out some of the main landmarks of the whole domain.

### APPRAISAL BY TESTS VERSUS APPRAISAL BY OBSERVATION IN NATURAL SITUATIONS

Attempts to appraise and describe a person can be grouped into two main categories: those that depend upon setting up special test situations and those that depend upon observing behavior in the actual naturally occurring situations of life. The usual earmarks of a test are that (1) it occurs at a specified time and place, (2) it consists of a set of tasks uniform for each person tested, and (3) it is seen as a test situation by the person being appraised. By contrast, evaluation based upon the naturally occurring situations of life is likely to (1) extend over an indefinite period, (2) be based upon situations that vary from person to person, and (3) not be perceived as a test by the person being appraised. The distinction between test situations and natural life situations is not an entirely sharp and clear-cut one, and we will have occasion to consider some *in-between* cases. However, it is usually clear whether we are dealing with a test as such or with observations under the natural conditions of life.

In thinking about the evaluation and measurement of man, we are likely to think primarily of tests narrowly defined, a test of arithmetic, a test of scholastic aptitude, or a test of auditory acuity. But we must remember that many of the important appraisals we make of people have always been, and will continue to be, based on observations of

them as they live from day to day. Appraisals of the nursery-school child's insecurity in relation to other children, of the 10-year-old's cooperativeness, or of the junior executive's initiative will almost necessarily be based upon observations of him over a period of time as he functions in his natural social group. Evaluations based on these observations have serious limitations. We are likely to find little uniformity from person to person in either the situations observed or the standards of judgment of the observers. But for some kinds of behavior we have no adequate tests to substitute for observations of natural situations—and very likely never will have.

Any complete picture of evaluation procedures must, therefore, pay attention both to test techniques and to devices for improving the observation of naturally occurring behavior. We will tend to prefer test situations where suitable ones can be devised. The examiner has more control over the situation, since he can present the same tasks or questions to everyone in the same way. He can usually get more precise results from a test and results that depend less upon the particular person making the appraisal. However, we must recognize that many significant aspects of individual behavior, by their very nature, defy reduction to a neat test. These can be appraised validly only as the individual functions in a natural life situation.

Of course, not all tests are perfectly frank and aboveboard. We shall have occasion to consider various types of test instruments in which the characteristics appraised are not those that the test seems to be getting at. Outstanding in this group are the so-called *projective tests* discussed in Chapter 15. What purports to be a test of "imagination" may in fact be directed at revealing anxieties, tensions, and inner emotional conflicts. Or a test of arithmetic computation may be rigged to yield a measure of cheating. But these are exceptions to the general rule that in a test the person knows that he is being tested and knows *what* is being tested.

## TWO FORMS OF TESTS

Within a defined test setting we may again recognize an important distinction, which depends upon whether the examinee leaves a permanent record of his behavior or whether it must be observed "on the wing" as it takes place. The first situation is represented by any test, such as one of reading comprehension, in which the examinee marks his answers on a paper. The marks are then permanently recorded and can be scored at leisure. The second type of test would be encountered in an appraisal of oral reading, for example, where errors

are noted by the listener as they occur or the quality is judged by the listener as the reader speaks.

In this comparison, again, the advantages with respect to reliability and objectivity usually fall on the side of the test that gives a permanent record, the test with answers on an answer sheet or a definite product produced. It is hard to observe and record behavior accurately as it is taking place. Inaccuracies and biases tend to creep in. The observer is hurried; his attention lapses. Consequently, in developing testing devices the tendency has been to make them of the sort that leaves a permanent record.

But young children cannot read or write, and many others are handicapped in a test that requires them to do so. Again, some types of performances, such as speaking or singing, are not readily reduced to a usable permanent record. It is also true that sometimes we are interested not merely in *what* a person does but also in *how* he does it. If a child gets the right answer for  $6 \times 7$ , does he get it quickly or slowly? Surely or with fumbling? By automatic habituation of the correct answer or by counting up from  $6 \times 6$ ? The process does not show in the written answer but can sometimes be observed if the child answers the problem orally or "thinks out loud."

There are test situations, therefore, in which we shall have to depend upon observations of the behavior as it takes place rather than upon scoring the written record. These test situations pose special problems. Observers must be taught what to look for. They must be taught what responses to record and how to record them. They must be trained in standards of judgment, so that the pronunciation that they accept as right, for example, will also be one accepted as right by other observers. It is for this type of test that special training of examiners is usually required.

## EXTERNAL OBSERVERS VERSUS SELF-OBSERVATION

As we move out of a test setting into observation of the individual's behavior in the natural situations of life, two distinct options are again open to us. We may rely upon some outsider to observe the person's behavior, someone such as his teacher, his employer, a friend, or a member of his family. Or we can ask him to report on his *own* characteristics as *he* sees them. These provide two quite different views of the individual, the one from the outside, the other from the inside.

The outside view is filtered through the biases and limited contacts of a particular outsider. The teacher, for example, sees only one

side of the youngster—the school side that is turned toward him. Furthermore, he sees it colored by his own prejudices and limitations. What he sees as “cooperation” may from another viewpoint appear to be docility; what he considers “insubordination” may appear to another to be independence.

The self-picture is limited by the reporter's lack of self-understanding and unwillingness to reveal himself to the watching world. We do not know ourselves perfectly. Some of our limitations, our petty meannesses and evasions, our weak and sensitive spots, we cannot face and admit even to ourselves. Still other shortcomings we recognize but are unwilling to acknowledge to an outsider.

Sometimes one set of limitations will seem more serious, sometimes the other. If a person is applying for a job he very much wants, we will probably feel that we can put more trust in the evaluations of outsiders than in his self-evaluation. He has too much at stake in the impression he makes. On the other hand, if he has come to us for help and guidance, his own more intimate self-picture may provide a better basis for counseling with him than will the impressions of an outsider. We shall need to become acquainted with evaluation instruments of both types.

## PLANNED VERSUS RETROSPECTIVE OBSERVATION

When we rely upon observations, either by the subject himself or by others observing him from outside, we may call for new observations made specially for us, or we may fall back upon the informal and undirected observations that have occurred in the past. Suppose we are studying the individual's tendency to become angry. We might ask him to keep track of all the times he got mad during the following week, noting down the circumstances for each anger episode, i.e., when it occurred, what precipitated it, what he did, etc. This would be an example of planned self-observation. By contrast, a second possibility would be to give him a list of situations that tend to annoy or irritate people. We might then ask him to look into himself and judge how readily he had tended to get angry at people who push in front in line, at being called by the wrong name, at being called down for something he did, and so forth. The self-observations would now be retrospective. If an outsider—say, a teacher—were doing the job, he might be asked to note down times during a specified period when he saw the particular pupil push, hit, or talk sharply to another. Or he might be asked to think back over his contacts with the child and

rate him on a scale ranging from "exceptionally calm and even-tempered" to "flares up and gets angry at the slightest provocation."

Again, there are advantages and disadvantages to both the planned and the retrospective type of observation. A major difficulty with systematic planned observations is that they are laborious and time-consuming. It takes a great deal of time and a high level of observer cooperation to get the necessary observations made. Partly because of this, the observations are likely to cover a limited time period and therefore to represent a rather meager sample of the individual's behavior. However, when observations are of actual current behavior, they tend to be more objective and less influenced by biases and the selective effects of memory than retrospective reports. The retrospective observations called for in self-report inventories and in rating scales have been widely used because of their administrative simplicity and because they summarize concisely the whole history of self-observation or contact with the person rated. But this type of summarizing judgment gives the biases of the respondent the fullest chance to express themselves.

### OBSERVATION AND TEST COMBINED—THE SITUATIONAL TEST

As we noted earlier, some behavior in test situations leaves no record behind but must be observed as it occurs. Here we have something of a hybrid involving both observation and test. The observer notes the specific errors a child makes when he reads aloud or his hesitations and false starts in spelling a word. Sometimes the "test" may involve a much more complex and total situation and more subtle types of behavior. In many of these "tests," the person being observed may not realize what is being observed (or even that he is being observed). So, if we want to appraise the individual's tendency to get angry, we may put him in a standard anger-producing situation. For example, we may give him a job to do and two intentionally stupid assistants who keep making mistakes and getting in the way. In so far as we are able to present each subject with the same task, we have a test situation. But we must depend upon the observations and judgments of outsiders to evaluate his behavior.

These complexly structured lifelike situations, which strive for the uniformity of a test situation and yet for the naturalness of real-life events, may be called *situational tests*. They represent a compromise between the objectivity and standardization of the testing approach and the naturalness of a real-life situation. This approach presents

interesting possibilities for getting at types of behavior that do not readily lend themselves to the conventional types of testing.

The practical problems faced in devising situational tests are very great. They call for elaborate staging if the naturalness of real life is to be preserved. In addition, the problems of obtaining satisfactory observations and adequate reports of them remain. For these reasons, situational tests have not been widely used. But they present an interesting type of tool, whose possibilities are only beginning to be explored.

### FUNCTIONS FOR WHICH MEASUREMENT HAS BEEN UNDERTAKEN

Broadly speaking, psychologists and educators have been interested in measuring in two general areas, what a person *can* do and what he *will* do. Measures of the first sort are measures of *ability*. In our discussion we will divide ability measures into measures of *aptitude* and measures of *achievement*. Again, roughly speaking, an aptitude test undertakes to measure what a person *could learn* to do, whereas an achievement test measures what he *has learned* to do.

The distinction between aptitude tests and achievement tests is far from a clear one, because we often use what a person has learned as a cue to what he can learn. Thus, a measure of the amount of knowledge of mechanical devices a person has gained in the past may be one of the most accurate indicators of the amount of further knowledge of things mechanical he will acquire in the future. The clearest distinction between aptitude and achievement tests lies in the direction of our interest. In an aptitude test, our interest is to predict what the individual *can learn* or develop into in the future; in the achievement test our interest is in what he *has learned* in the past.

Measures of the second major category—of what the person *will* do—correspond to the area we may roughly label *personality* measurement. This is a somewhat broad and loose definition of personality. It is also a somewhat external one. That is, we have indicated a concern for what a person *does* rather than for how he feels or what his inner urges and conflicts are. We may be interested in those to a degree. But, so far as a testing or observational procedure is concerned, it is always based on what a person does—how he acts, what answers he marks, or what he says. His actions are the basic material that we study.

In the long run, his future actions are what we want to predict: whether he will graduate from college, whether he will continue in

and apply himself to a clerical type of job, whether he will behave in a more socially acceptable fashion after a particular type of therapy. We may perhaps make these *predictions* more surely if we organize the test and observational appraisals around certain concepts of interests, needs, or conflicts. But these terms describing the inner life of the individual represent inferences that we make as a way of structuring and organizing the observations of the individual's behavior. We cannot see a need for approval. What we observe is that a child brings things into class, attempts to talk at all times, buys candy for other children, and tries to join any social group in the playground. We may *infer* a need for approval as an underlying factor related to the various behaviors.

When we try to measure what a person *will* do, as distinct from what he *can* do, we encounter some special problems. These are primarily problems of intentional distortion of the test results. In an ability test we want each individual to try hard and do the best he can. But in personality measures, we do not want to know how cooperatively a person can behave or how energetic he can be. We want to know to what degree he *typically* does show energy or behave in a cooperative manner. In a limited test situation, where the nature of the test is clear to the examinee, everyone can put his best foot forward. He can probably muster up all the virtues for a special occasion. But will he in other situations? It is this question, the question of the degree to which behavior in an identifiable test situation will represent behavior in real life, that pushes us into disguised tests and into observational evaluations of personality characteristics.

#### ASPECTS OF PERSONALITY

It will be convenient to use a number of terms to refer to certain fractions or aspects of personality that we may wish to evaluate. These terms and the meanings that attach to them are discussed briefly below.

*Character.* Character traits are aspects of individual behavior to which a definite social value has been attached. *Honesty, cooperativeness, thrift, kindness, and loyalty* are all labels for social virtues. Educational and religious organizations have always been concerned with the inculcation of such virtues. Based on this concern there have been developed a number of evaluation procedures that we shall refer to as measures of character.

*Adjustment.* Educators and psychologists have long been concerned with the concept of adjustment. The mental hygiene approach as applied both in and out of school has striven to develop "well-



adjusted personalities." Maladjustment is recognized in individuals who fail to fit into the social group or who appear to live unhappy and unproductive lives. As with character, degree of adjustment represents a social judgment, and what is conceived to be well-adjusted behavior varies from one culture to another, depending upon what is normal for that culture. Normal behavior in our competitive, acquisitive society might seem pathological if transferred to a South Sea island. Adjustment will mean, then, behavior patterns that enable the person to get along in and be comfortable in his social setting—typically, the setting of middle-class, twentieth-century American-European culture. We shall encounter a group of instruments designed to evaluate deviations from this norm—the tendency to show maladjusted behavior or behavior typical of people who do not get on happily and successfully in our culture.

*Temperament.* From early days observers of human nature have noted conspicuous differences in energy level, prevailing mood, and general style of life. Literary men and men of science alike have proposed systems for classifying temperaments. Hippocrates, for example, proposed that men could be divided into the sanguine (energetic and cheerful), choleric (energetic and irascible), phlegmatic (sluggish and placid), and melancholic (sluggish and sad), and proposed physiological bases for these distinctions. There have been many other classifications before and since. Appraisals of such dimensions as these we shall speak of as measures of temperament.

*Interest.* The individual makes a variety of choices with respect to the activities in which he engages. He shows preferences for some, aversion to others. Appraising these tendencies to seek or avoid particular activities constitutes the domain of interest measurement.

*Attitude.* The individual responds with enthusiasm and aversion not only toward activities but also to social groups, social institutions, and the other aspects of his world. These reactions, with their various ramifications, constitute the individual's constellation of attitudes. Various devices have been developed for evaluating these prejudices pro and con, and these constitute the field of attitude measurement.

## CONCLUDING STATEMENT

In summary, then, approaches to the measurement of the individual cover a great diversity both of methods and of content areas. Variations of method may be represented by the following outline:

1. *Test methods*, involving a defined task and testing period.
  - A. Permanent record or product available for scoring or analysis.
  - B. Process must be observed and evaluated as it occurs.

- II. *Observational methods*, in which behavior is observed in the natural situations of life.
  - A. *Self-observation*, in which the individual reports on his own reactions, as far as he is aware of them.
    1. Planned observations, planned in advance to cover a specified period.
    2. Retrospective observation, based on present memory and evaluation of past reactions.
  - B. *Observation by an outsider*, in which relative, employer, teacher, etc., reports on the individual's reactions.
    1. Planned observations.
    2. Retrospective observations.
- III. *Mixed methods*, characterized by some of the aspects of a test but also relying upon observation and evaluation of observed behavior.

Advantages and problems of these approaches have been sketched in but will need to be considered in more detail as specific methods are elaborated in later chapters.

Aspects of the individual for which evaluation procedures have been developed and in which we shall be interested include the following:

- I. *Abilities*, evidences of what the individual can do if he tries.
  - A. *Aptitudes*, performances serving as indicators of what he can learn to do.
  - B. *Achievements*, performances used to show what he has already learned to do.
- II. *Personality variables*, indications of what an individual will do, of how he will respond to the events and pressures of life.
  - A. *Character*, certain qualities defined by society as estimable or the reverse.
  - B. *Adjustment*, degree of ability to fit into and live happily in the culture in which one is placed.
  - C. *Temperament*, qualities relating to energy level, mood, and style of life.
  - D. *Interests*, activities that are sought or avoided.
  - E. *Attitudes*, reactions for or against the people, the phenomena, and the concepts that make up society.

This analysis of aspects of the individual is neither complete nor detailed. However, it serves to indicate the range of measures with which we shall be concerned in the following chapters.

## QUESTIONS FOR DISCUSSION

1. It would generally be agreed that personality measures are less satisfactory than measures of aptitude or achievement. What factors give rise to this?
2. How would you fit each of the following into the classification of measurement methods given in the chapter?

- a. Anecdotal records kept by a teacher, describing behavior in his classroom.
  - b. An autobiography written by a pupil for a high-school counselor.
  - c. An individual intelligence test in which both questions and answers are given orally.
  - d. A Boy Scout's record of "good deeds," kept over a 2-week period and reported to his Scoutmaster.
3. Illustrate, from your reading or experience, each of the categories of measurement methods in the outline on pp. 24-25.
4. How would you fit each of the following into the outline of aspects of the individual to be evaluated, given on p. 25?
- a. Observations of how well a high-school student gets along with adults.
  - b. A pupil's expression of his preferences for books in an annotated list of titles.
  - c. A kindergarten child's performance on a test of readiness to learn reading.
  - d. A pupil's performance on an English test, used to place him in the appropriate section.
  - e. Ratings of a pupil on his loyalty to his friends.
5. From your reading or personal experience, give an illustration of measurement procedures for each of the aspects of the individual identified in the outline on p. 25.
6. A class has just finished a unit on etiquette, and the teacher wishes to evaluate the effectiveness of the unit. Which of the methods outlined on pp. 24-25 might she use? What would be the advantages and limitations of each?

## Chapter 3



# The Teacher's Own Tests

In this book dealing with educational and psychological measurement procedures, we have elected to start with a consideration of the teacher's own tests. We have done this for several reasons. In the first place, informal test making is an operation that is familiar to every teacher, and the outcomes of such test making are familiar to every student. In the second place, because the teacher-made test is so widely used and has such an important place in evaluating student achievement, it strongly influences students' views toward tests and test-taking specifically and toward education generally. In the third place, the techniques of testing available to every teacher form the backbone of standardized tests of achievement and of aptitude. Furthermore, the quality of the items on a standardized test and the adequacy of the coverage of a standardized test are judged by precisely the same standards that apply to teacher-made tests.

### THE ROLE OF TEACHER-MADE TESTS

Evaluation of pupil progress is a major aspect of the teacher's job. A good picture of where the pupil is and of how he is progressing is fundamental to effective teaching by the teacher and to effective learning by the pupil. The evaluation \* procedures the teacher uses with his group serve a number of functions. We will identify four, commenting briefly upon each of them. All the procedures the teacher develops for pupil evaluation may serve these functions, but we shall be concerned to point out how they may be served by the more formal evaluation instruments called tests.

\* The term "evaluation" as we use it here is closely related to measurement. It is in some respects more inclusive, including informal and intuitive judgments of pupil progress. It also includes more definitely the aspect of valuing—of saying what is desirable and good. Good measurement techniques provide the solid foundation for sound evaluation, whether of a single pupil or of a total curriculum.

### MOTIVATION

To some degree, varying from pupil to pupil and from class to class, tests determine when students study, what they study, and how they study. Tests that are well constructed and effectively used can motivate students to develop good study habits, to correct errors, and to direct their activities toward the achievement of desired goals. Tests that are poorly constructed or used punitively can just as effectively discourage the students or misdirect their learning. Testing procedures control the learning process to a greater degree, perhaps, than any other teaching device.

### DIAGNOSIS AND INSTRUCTION

Testing serves to diagnose weaknesses and to provide practice for available knowledges and skills. The items on which an individual fails or on which many members of a class group fail can serve to identify points needing further study whenever the test task is sufficiently precise for the nature of the failure to be identified. The function of a test as a rehearsal of knowledge and a guide for further study has long been recognized.

### DEFINING TEACHING OBJECTIVES

What a teacher emphasizes in his evaluation of pupils, and particularly in the more formal evaluation represented by tests, defines to his students what that teacher considers important. This definition is presented in a much more forceful way than any pretty speeches that the teacher may make. The teacher may avow, to his students or to his colleagues, that he considers the ability to apply facts to real situations and to understand basic principles to be much more important than just learning facts. But if his tests ask only for names, dates, places, and sentences from the book or his lectures, those will be his functional objectives, and those will be the things that his students will study—the docile ones who are influenced by him anyhow. We may know a teacher by the tests he makes. They tell what he is truly valuing in his pupils, even though he himself does not know it, and they influence profoundly what his students will learn.

### DIFFERENTIATION AND CERTIFICATION OF PUPILS

The teacher inevitably has a responsibility for certifying pupils' accomplishments to higher levels of the educational enterprise and to the world outside the school. The testing procedures he uses help

him to arrive at the judgment that is recorded in his mark, letter of recommendation, or other evidence of approbation or disapprobation.

In view of the many functions they serve and in view of the disservice that may be done the pupil from poorly conceived or executed evaluation instruments, it is important that the teacher's evaluation devices be well thought out and well made. To evaluate the range of outcomes in which a modern school is interested—understanding as well as knowledge, appreciation as well as skill, ability to apply as well as to reproduce, attitudes and interests as well as achievements—the teacher must call upon a variety of types of appraisal. He must profit from observation of classroom performance by recitation, by participation in informal discussions, by contribution to group enterprises. He must size up the student in conference, interview, and informal discussion. He may have occasion to rate the products produced in laboratory or shop and to appraise the quality of assignments carried out outside of school. He will also almost certainly make some use of class tests. Some of the objectives of his teaching can be measured efficiently, realistically, and completely by pencil-and-paper tests. Some can be measured only partially by such means. Some cannot be measured at all in this way. This chapter and the next are concerned primarily with those objectives that can be measured with tests and with the improvement of testing procedures to measure them. Some consideration will be given to observational procedures, ratings, and other types of appraisal devices in later chapters.

## PLANNING THE TEST

The primary function of any evaluation procedure is to determine to what extent students have achieved the objectives of instruction. If a test is to serve this function effectively, it must be planned with that end in view. A test which "just grew" is unlikely to correspond very well to the teacher's stated objectives. This is particularly true in the case of objective tests, and it is here that careful planning is especially important. However, one should not overlook the importance of a good test plan even in the case of an essay test.

If the teacher just sits down and writes objective test items, the test is likely to be out of balance. It is easier to write simple factual items than it is to write items that call for understanding of generalizations or application of principles. It is easier to write items on some topics than on others. As a result, the teacher is likely to end

up with an overload of items calling for simple information about the more testable topics. The same thing is true, to a degree, of essay tests. The outcomes measured by the test will then show a poor correspondence with those espoused by the teacher. What the pupils emphasize in their learning will soon follow what they find is emphasized in their tests, and the tests will fail to foster the learnings in which the instructor is most interested.

### DEFINING OBJECTIVES

The thoughtful planning of a test involves several steps. The first and most important step is to define the objectives that are to be appraised. Before he can evaluate whether a student has achieved the objectives of instruction, a teacher must be able to state what the student was supposed to have achieved. Moreover, objectives that are to be evaluated must be stated in terms of pupil behavior. We must be able to specify the *processes* or *activities* that a student is expected to display if he has achieved the objectives. What do we expect him to know? What kinds of applications do we expect him to be able to make? How do we expect him to think or to solve problems? What actions on his part will show that he has acquired the attitudes that we are trying to inculcate? In other words what things must a student *do* to show that he has acquired the knowledge, understandings, skills, attitudes and appreciations that we say we have been trying to teach.

The failure to define objectives in terms of student behavior probably accounts for much of the inadequacy in evaluation of student progress in schools and also for the very poor quality of many classroom tests. Defining the objectives of instruction in terms of pupil behavior is not an easy task but it is necessary before a good test can be constructed or effective evaluation can be done.

The real work of defining objectives must usually be done by the teacher himself, perhaps assisted by his colleagues, and working from his textbook or course outline. In many schools the teacher has available a curriculum guide or course of study which does contain a set of objectives. But objectives listed in these sources tend to be too vague and global to be useful as a guide for evaluation. They need to be broken down into more specific components if they are to provide a sufficiently exact definition of just what the broad, global objectives mean.

Let us look at an actual example. In the section below are listed the objectives stated as the desired outcomes for an eighth-grade social-studies unit on the functioning of our national government.

**Objectives of a Unit on How Our National Government Functions**

1. Has a basic foundation of facts and information necessary to an understanding of the unit.
2. Understands why the Declaration of Independence was written.
3. Understands the ideas embodied in the Declaration of Independence.
4. Understands the Articles of Confederation.
5. Understands Articles I through VII of the Constitution.
6. Understands the Bill of Rights and other amendments to the Constitution.
7. Can use and interpret maps.
8. Can locate and interpret data.
9. Can do critical thinking.
10. Derives personal satisfaction from social studies reading.
11. Is able to plan, execute, and evaluate committee projects.
12. Uses parliamentary procedures.
13. Develops a love for and loyalty to the principles of the government of the United States.
14. Develops an abiding interest in civic affairs beyond the years of formal education.
15. Has an appreciation for the principles of a democracy which formed the basis of our government.

As stated, these objectives embody many of the faults typically found in the statements of objectives available to teachers in courses of study. Some of these may be pointed out and illustrated.

1. *The Objective Is One That Cannot Be Achieved, and Certainly Cannot Be Evaluated within the Unit.* Objective 14 refers to adult life, and not to anything that characterizes the pupil in the eighth grade. It is an expression of a pious and worthy hope, but of little help in guiding the teacher as to what he should do with a pupil or look for in him.

2. *The Objective Is Expressed in Terms of Unobservables.* Objectives 13 and 15 state worthwhile hopes, but provide no guidance to the teacher as to what the eighth grader is to do to show his love, loyalty, and appreciation. How does a student exhibit appreciation for the principles of democracy? Through giving lip service to the words and symbols? Through accepting individuals who differ from himself in various ways? Through participating in school government? Behaviorally, what do these objectives mean?

3. *The Objective Bears Little or No Relationship to the Content of the Unit.* Objective 7 is one that has little relationship to the unit being studied. There are certainly many better units in which to build up map-reading skills. Such skills play very minor roles, if any, in this particular content.



4. *The Statement of the Objective Implies a Process Quite Different from the One That Is Taught.* Objectives 2 through 6 start with the word "understand." But what does the word really mean in this context? Consider 2, for example. Further study of the curriculum guide brings out that the pupil is to "understand" that the Declaration of Independence was written to explain America's cause in the War of Independence. However, this point is specifically and explicitly made in the textbooks. A pupil could produce this statement on an examination on the basis of direct recall of what he had learned. No real understanding is called for. The same is true of objectives 3 through 6. These objectives should more appropriately begin with "knows," "can recall," or "can state," because these words more accurately reflect the process of reproducing from memory.

The term "appreciation" in objective 15 is also often used quite loosely. It is often used to refer to information about something, rather than an affective or aesthetic reaction to it. Thus, again, it may be more realistic to talk about knowledge of the principles, and perhaps ability to interpret or apply them, than about "appreciation" of them.

5. *The Objective May Be Inappropriate to the Level of Instruction.* To be realistic, we can expect eighth graders to achieve only a very limited understanding of this particular unit. The teacher should recognize this, and place "understanding" in a proper perspective both in the weight that is given to it and the level of sophistication that is expected.

The teacher *can* both develop and test for the level of understanding appropriate at a given grade level if he can identify the processes on the part of the student that represent understanding. The student can show understanding at different levels by (1) expressing concepts and principles in his own words, (2) pointing out similarities and differences not explicitly pointed out in class or text, (3) pointing out relationships, and (4) applying his information to situations about which he has not been taught.

We could point out other specific defects in the list of objectives as they are stated, but let us stop and see how the list *might* be revised so that the objectives would provide a better guide both for teaching and evaluation. For this purpose we want to separate the process from the content. A good guide for setting up objectives that relate to process rather than content is given by Bloom et al.<sup>1</sup> A revised set of process objectives for this unit is given below. When the word "knows" is used, it means the ability to reproduce, recall, recognize or define.

1. Knows terms and vocabulary.
2. Knows dates, events, persons, and places.
3. Knows generalizations, concepts, and principles.
4. Can trace the sequence of historical development (of our national form of government.)
5. Can express generalizations and concepts in his own words.
6. Can point out relationships, similarities, and differences.
7. Can apply generalizations and principles to particular concrete situations that are new to him.
8. Can use parliamentary procedures in committee or class meetings.
9. Can plan, execute, and evaluate committee projects.
10. Shows support either orally or in writing for governmental actions that follow the democratic principles on which the government is based.
11. Expresses concern either orally or in writing for an individual or group being deprived of the rights guaranteed by our constitution.

This set of objectives keeps the basic intent of the original set of objectives. For example, objectives 1 through 7 in the revised list include the intent of objectives 1 through 6, 8, 9, and 15 on the original list of objectives. Objectives 8 and 9 in the revised list parallel 11 and 12 on the original list. Objectives 10 and 11 on the revised list are redefinitions of 13 on the original list. Those objectives on the original list that were totally unrealistic or did not apply to the unit have been eliminated in the revised list.

### OUTLINING CONTENT

The second step in planning for a test is to outline the content to be covered. The outline of content is important because the content is the actual vehicle through which the process objectives are to be achieved. An outline of content for our illustrative example follows.

#### *Outline of Content of a Unit on How Our National Government Functions*

- I. The Foundations of Our Constitutional Government (Time allotment: 2 weeks)
  - A. Early English documents—the Magna Charta, the Petition of Right, and the Bill of Rights
    1. Provisions of these documents
    2. Their influence on our heritage
  - B. Mayflower Compact, Fundamental Orders of Connecticut, The New England Confederation, The Albany Plan
    1. Principles embodied in these documents
    2. Influences of these documents
  - C. Representative Assemblies in the Colonies

- D. Declaration of Independence
  - 1. Events leading up to the drafting of the Declaration of Independence
  - 2. Drafting the document; ideas embodied in the declaration
    - a. Principles of government the delegates believed in
    - b. List of grievances against King George III
    - c. A declaration of freedom
  - 3. Adoption of the Declaration of Independence on July 4, 1776
- E. The Articles of Confederation (1781-1789)
  - 1. Provisions
  - 2. Weaknesses
  - 3. Constitutional convention called by Congress in 1787
- F. The Constitutional Convention—1787
  - 1. Members
  - 2. Stated purpose
  - 3. Need for strong national government
  - 4. Disagreements solved by debate and compromise
  - 5. Contributions of influential men at the convention
  - 6. Ratification of the Constitution
- II. Principles and Development of Our Constitutional Government  
(Time allotment: 4 to 6 weeks)
  - A. Purpose as stated in the preamble
  - B. Federalism versus Confederation
    - 1. Division of powers between central government and states
    - 2. Advantage of federalism
  - C. Unitary and Federal Government
  - D. Division of power in our federal system
    - 1. Reasons for separation of powers
    - 2. Division of powers in federal government
    - 3. Division of powers among local units
    - 4. Limitations placed on federal government
  - E. The Three Branches and the System of Checks and Balances
    - 1. The executive, legislative, and judicial branches
    - 2. Meaning of checks and balances
  - F. The Articles of the Constitution
    - 1. Provisions of the articles
  - G. The development of the Constitution
    - 1. Elastic clause
    - 2. Changing the Constitution
      - a. Amendments—process
      - b. Statutes
      - c. Court decisions
      - d. Customs and usage
  - H. The Bill of Rights

#### PREPARING THE TEST BLUEPRINT

The content outline and a statement of process objectives represent the two dimensions into which a test plan should be fitted. These two dimensions need to be put together to give a complete framework and

to see which objectives *relate especially* to which segments of content. In planning for the *total evaluation* of a unit the teacher would be well advised to make a *blueprint* covering *all* objectives and add a column at the extreme right indicating the method or methods to be used in *evaluating student progress toward* achieving the objectives. However in making a blueprint for a *test*, only those objectives that can be measured either wholly or in part by a paper-and-pencil test should be included.

In the list of revised objectives, objectives 8 through 11 cannot be measured by a paper-and-pencil test. For example, objective 8, "Can use parliamentary procedures in committee or class meetings," can be evaluated only by putting the student in a somewhat *formal meeting* and observing whether he follows parliamentary procedures. On a paper-and-pencil test the teacher could determine whether a student *knows* parliamentary procedure but not whether he *uses* it. Since a student cannot use parliamentary procedure unless he knows it, such testing of knowledge is sometimes worthwhile, but the teacher should remember that in this unit the objective was stated in terms of using rather than of knowing. Observation of performance in assigned class groups and in *informal school activities* would seem promising approaches to evaluating objectives 9 through 11. The teacher should remember that no single test or evaluation medium can measure all the objectives that he is trying to achieve.

Figure 3.1 shows the process objectives in our revised list that can be measured by a paper-and-pencil test and their relation to the content of the unit. The objectives are listed in the left-hand column of the chart and the titles of the two major content areas are used to head the two right-hand columns. Each cell is then filled with notes suggesting terms, dates, events, generalizations, relationships, or applications that a teacher might consider important specific examples of that content and that process. For example, the upper left hand cell contains four terms—*inalienable rights*, *tyranny*, *compromise*, and *confederation*—that this teacher considered it important for students to know. The cell below that contains certain events, dates, and documents that this teacher considered important. The other cells have been filled in the same way in order to specify more precisely the content to be included and what the student is supposed to be able to do with that content.

The preparation of such a two-dimensional outline is undoubtedly an exacting and time-consuming task. The busy classroom teacher may often fall short of achieving such a complete analysis. There is no question, however, that attempting the analysis will go far toward clari-

Questions	1. The foundations of our Constitutional Government (30% of all items)	II. Principles and Development of our Constitutional Government (70% of all items)
1. Name three or four principles and axioms (10%)	1. Unalienable Rights 2. Economy 3. Consensus 4. Confederation (1 or 2 items)	<ul style="list-style-type: none"> <li>Preamble</li> <li>Legislative</li> <li>Executive</li> <li>Judicial</li> <li>Federalism</li> <li>Unitary governments</li> <li>Supremacy</li> <li>Executive checks</li> </ul> (4 items)
2. Name three or four principles and axioms (5%)	<ul style="list-style-type: none"> <li>Articles of Confederation - 1781</li> <li>Magna Carta - 1215</li> <li>Bill of Rights - 1689</li> <li>Declaration of Independence - 1776</li> <li>Declaration of Constitutional Convention - 1787</li> <li>Articles of Confederation</li> <li>Constitution of 1787</li> </ul> (10-11 items)	<ul style="list-style-type: none"> <li>Ratification of Constitution - 1788</li> <li>Substantive powers - (2 items)</li> </ul>
3. Name three or four principles and axioms (15%)	<ul style="list-style-type: none"> <li>Preamble of Magna Carta, Bill of Rights</li> <li>Articles of Confederation</li> <li>Principles embodied in the Declaration of Independence</li> <li>Articles of Confederation</li> <li>Articles of Confederation</li> <li>Articles of Confederation</li> <li>Articles of Confederation</li> </ul> (16 or 17 items)	<ul style="list-style-type: none"> <li>Articles on the states and National Government</li> <li>Articles belonging only to the states</li> <li>Articles belonging only to the National Government</li> <li>Articles shared by states and National Government</li> <li>Articles of Joint III. Articles</li> <li>Changes and amendments to Constitution - how made</li> <li>Changes of Bill of Rights</li> <li>Changes of Bill of Rights</li> </ul> (14 or 15 items)
4. Can trace development of our national form of government (20%)	<ul style="list-style-type: none"> <li>Identify influences and trends that contributed to the establishment of representative assembly in the colonies</li> <li>Explain the development of representative government in the colonies between 1607 and 1776</li> </ul>	<ul style="list-style-type: none"> <li>Identify events, forces and laws that led to separation of powers in federal government</li> <li>Identify forces that led to amendments to the Constitution (3 or 4 items)</li> </ul>

fyng the objectives of a particular unit and toward guiding not only the preparation of a sound test but also the teaching of the unit itself.

Once the basic outline has been prepared, the test maker must decide upon the relative emphasis to be given to the several content areas and process objectives. The number of questions that can be presented in a test is limited by the testing time available and by the ability and background of the students to be tested. Since time does not permit him to include everything, the test maker must select a sample of questions. The sample should truly represent the emphasis given in his teaching, both with respect to content and with respect to the process objectives. This can be done by having the proportion of questions in each content area correspond to the proportionate emphasis given to that topic, and the proportion of items calling for each process correspond to the importance the teacher considers that process to have in the learnings that the pupils are to achieve. The decisions made by the teacher in dividing up the questions on a test are necessarily subjective ones. The basic principle underlying these decisions is that the test should maintain the same balance in relative emphasis on both content and mental processes that the teacher has been trying to achieve through his instruction. This allocation of differing numbers of items to different topics and process objectives is one way of *weighting* these topics and objectives differentially in the test.

In the illustration of Figure 3.1, the test maker decided that about 30 per cent of the items should be on Topic I and 70 per cent on Topic II. This corresponds roughly to the allocation of teaching time to the two topics given on pp. 33-34, i.e., 2 weeks to Topic I and 4 to 6 weeks to Topic II. Time spent on the topic and space given to it in the textbook can help to guide the teacher's judgment of the basic importance of the topic and weight to be given it.

The test maker in Fig. 3.1 allocated 35 per cent of the total number of items to objective 3, 25 per cent to objective 4, 10 per cent each to objectives 1, 5, 6, and 7, and 5 per cent to objective 2. These reflect a judgment by this test maker that objective 3, knowledge of concepts, generalizations and principles, is clearly the most important objective of the unit, that objective 4, ability to trace sequences of historical events, is next most important, and that the others are of about equal, but lesser importance. Names and dates (objective 2) are relegated to a minor role. Note that here, as in the topical outline, 100 per cent has been distributed among the different categories.

The test maker must also decide now whether he will use essay test questions or short-answer, objective items, and if he decides on objective items he must decide which type or types he will use. The

choice is governed, at least in part, by the objectives to be measured. The next section of this chapter will provide some discussion of the advantages and disadvantages of essay questions, in relation to the objective type of item. Different types of objective items will be described and compared in Chapter 4.

At about this point the total number of essay questions or objective test items must be decided upon. This is primarily a function of the time available for the test and the type of items being used. Different types of objective items differ in the time allotments they require, and, of course, an essay question demands a great deal more time than an objective item. It is almost impossible to state in general terms how much time should be allowed per item for objective items of a specific type. The appropriate time allowance is affected by a host of different factors. Among the most important are (1) the age of the pupils being tested, (2) the length and complexity of the item, (3) the type of objective being tested—knowledge of fact or concept versus application to new situation, (4) the amount of computation, if any, required by the item, and (5) the relative interest of the examiner in speed versus power—the amount the pupil can do with unlimited time.

In general, it seems undesirable to emphasize speed in an achievement test designed to measure one's range of information or ability to apply knowledge. Most teacher-made tests should be power tests; i.e., there should be enough time so that at least 80 per cent of the students can attempt to answer each item. As a teacher becomes familiar with the kinds of students he usually has in a class, he will be able to judge the number of items he can include in a given amount of testing time while still having a power test. As a rough rule of thumb, the typical student might require from 30 to 45 seconds to read and attempt to answer a simple factual type multiple-choice or true-false item, from 75 to 100 seconds to read and attempt a fairly complex item requiring problem-solving or some computation. If the test items are based on a reading passage, tabular material, map, or graph, time must be allowed for reading and examining the material.

Adequately to sample achievement in a large segment of work, i.e., the content of a whole semester, may require more items than can reasonably be included in a single-period test. The only satisfactory solution to this problem is to allow two or more periods for testing. If a single unit of sufficient length is unavailable or seems likely to go beyond the attention span of the group, the natural solution is to break the test up into two or more subtests that can be given on successive days.

Once the teacher has decided upon the total number of items to be included in the test, he should go back to the blueprint and determine how many items are needed for each cell. In the sample blueprint, Fig. 3.1, the total time available for testing was 50 minutes. The test maker decided to have a total of 60 questions on the test. Applying the percentages in the blueprint, approximately 18 questions should be on Topic I (30 per cent) and 42 questions should be on Topic II (70 per cent). The 18 questions on Topic I are distributed in the cells of that column according to the weights assigned to the objectives. To obtain the number of items for each cell, one multiplies the number of items for Topic I (18 items) by the percentage assigned to the objective in each row. For example, to determine the number of items for the first cell in Topic I, we multiply 18 by 0.1 (10 per cent) which gives 1.8 items. Since this product is between 1 and 2, we can note that we should have either one or two items covering this content and this objective. The other cells in the blueprint are filled in by the same process. It is probably desirable to indicate a range of frequency for each cell, as was done in our example, in order to provide flexibility if difficulty is encountered in writing acceptable items for certain cells. The frequencies are to be thought of as a guide and not as a strait jacket.

After all the items for the test have been constructed, the teacher should make a final check by sorting the items in piles to match the blueprint in order to make sure that the two agree.

The above discussion of allocation of items to topics and objectives applies primarily to objective tests made up of a large number of items. The same degree of analysis hardly applies to an essay examination, which will at best be composed of a relatively small number of items. But these few items should also be distributed over the content and the process objectives so that the test represents as well as possible the explicit goals of instruction.

A final decision that comes in as part of the preliminary planning concerns the desired difficulty of the test items. The decision depends in part upon the purpose of the test. When the test is to measure *mastery* of the basic essentials in an area, the questions should be limited to basic essentials. If the unit has been well taught, all the items may then turn out to be very easy for the group. When the purpose of the test is to *discriminate* levels of achievement of different members of a group, i.e., to serve as a basis for ranking or grading, some items should be very easy, most of them should be of moderate difficulty, and a few should be difficult enough to spread out the ablest



members of the group. *Difficulty*, in this context, is defined in terms of the percentage of examinees who get the item right.

Our test plan now consists of:

1. An outline of content and objectives.
2. Specific suggestions of what might be covered under each combination of content and objective.
3. An allocation of per cents of the total test by content area and by objective and an estimate of the total number of items.
4. Specifications for the spread of item difficulties.

The next task is to prepare the actual test items. In the remainder of this chapter we will discuss the choice of item types and guides for improving the writing of essay questions. In Chapter 4 we will discuss guides for improving objective-type items.

## ADVANTAGES AND LIMITATIONS OF ESSAY AND OBJECTIVE TESTS

Teacher-made tests may be divided into two broad categories, essay or free-answer tests and objective tests. One hears many arguments about whether essay tests or objective tests should be used in schools but these "either-or" arguments are pointless. Neither the essay test nor the objective test is satisfactory as the sole type of test to measure academic achievement. Each type has its own advantages and limitations and each has its place. The problem is to use each type of test in those situations where its advantages are maximized and its weaknesses minimized.

### THE ESSAY TEST

The essay test consists of such problems as:

Compare the organization and powers of the central government under the Articles of Confederation with the organization and powers of the central government under the Constitution.

Why did the merchants and business men particularly desire to have the Articles of Confederation changed?

The Fifth Amendment to the United States Constitution states that *no person shall be deprived of life, liberty, or property without due process of law*. In your own words, explain what the underlined part of the statement means.

Why is the Magna Charta considered to be an important milestone in the establishment of a democratic government?

The essential characteristics of the task set by an essay test are that each student

1. Organizes his own answers, with a minimum of constraint.
2. Uses his own words (usually his own handwriting).
3. Answers a small number of questions.
4. Produces answers having all degrees of completeness and accuracy.

In these characteristics lie both the strengths and weaknesses of the essay examination. Let us consider each in turn.

*The Student Organizes His Own Answers.* Herein lies the distinctive advantage of the essay examination. It requires the student to produce, rather than merely to recognize, the answer. Thus, it minimizes the possibility of getting the answer by blind guessing or by using little cues to outguess the test maker. It can, if the questions are well prepared, bring out the examinee's ability to select important facts or ideas, relate them to one another, and organize them into a coherent whole. Emphasizing this integrative type of product, it elicits, so it is claimed, better study habits in those who are preparing for it.

*The Answer Is in the Student's Words and Handwriting.* At this point a premium is placed upon verbal fluency and skill of expression. The student who is able to write effectively will often get a higher grade than another student who clothes the same ideas in less attractive garb. Too often verbal fluency and aggressive salesmanship, bluffing, in short, pass for knowledge of the subject. In addition to skill in writing, quality of handwriting frequently influences the grade on an essay test. How often has a student been penalized because the instructor became irritated by poor handwriting, or could not be bothered to decipher obscure "hen tracks"? Effective written expression and good penmanship may be legitimate objectives of the educational enterprise, but they should be evaluated in their own right. They should not be allowed to contaminate our appraisal of a student's understanding of the causes of Hitler's rise to power or of Newton's laws of motion.

*The Test is Limited to a Small Number of Questions.* When the individual must organize and compose an answer of some length, as with questions like those on p. 41, the number of questions is inevitably limited. The time required to answer a single question makes it impossible to include more than five or ten questions in even a fairly lengthy test. This tends to result in what we might call a "lumpy"

sampling of what the student knows. We sink four or five big shafts into the mine of knowledge that the student possesses. If these happen to hit pay dirt, the student does well; but if they hit the gaps in his knowledge, he does poorly. With this small number of samples, chance is likely to play a relatively large part. We may get a very unfair sample of a particular student's knowledge.

Of course, it is possible to ask free-response questions that call for quite short answers. We might ask: What qualifications does the Constitution set for United States Senators? This question requires only a list of qualifications or a sentence or two for a complete answer. Questions such as this are transitional between the essay and objective test. They can be numerous and can sample many items of knowledge or understanding. However, they sacrifice the main feature of the essay question—the requirement that the examinee put together an organized answer in which he relates, evaluates, and integrates a number of facts and ideas.

*Answers Are of All Degrees of Correctness.* The bugaboo of the essay examination is the laborious and subjective operation of evaluating the answers. That it is laborious any teacher who has ever graded a set of essay papers for even a middle-sized class can testify. That the grading is subjective and relatively undependable has been shown by a number of separate studies.

Consider the following answers written by two eighth-grade students to the question "Compare the powers and organization of the central government under the Articles of Confederation with the powers and organization of our own central government today."

#### *Student A*

Our government today has a president, a house of representatives, and a senate. Each state has two senators but the number of representatives is different for each state. This is because of compromise at the Constitutional Convention. The Articles of Confederation had only a Congress and each state had delegates in it and had one vote. This Congress couldn't do much of anything because all the states had to say it was alright. Back then Congress couldn't make people obey the law and there wasn't no supreme court to make people obey the law. The Articles of Confederation let Congress declare war, make treaties, and borrow money and Congress can do these things today. But Congress then really didn't have any power, it had to ask the states for everything. Today Congress can tell the states what to do and tax people to raise money they don't have to ask the states to give them money. Once each state could print its own money if it wanted to but today only the U. S. Mint can make money.

*Student B*

There is a very unique difference between the Central Government under the Articles of Confederation and the National Government of today. The Confederation could not tax directly where as the National Government can. The government of today has three different bodies—Legislative, Judicial, and Executive branches. The Confederation had only one branch which had limited powers. The confederate government could not tax the states directly or an individual either. The government of today, however, has the power to tax anyone directly and if they don't respond, the government has the right to put this person in jail until they are willing to pay the taxes. The confederation government was not run nearly as efficiently as the government of today. While they could pass laws (providing most of the states voted with them) the confederate government could not enforce these laws, (something which the present day can and does do) they could only hope and urge the states to enforce the laws.

These two answers together with three other answers written by students in the same class were given to two groups of graduate students in courses in measurement or evaluation. Both groups of students were provided with a model answer to the question and given the following instructions:

Instructions: The essay question was a part of a social studies test consisting of fifty objective items and one essay question. The students were given 25 minutes to write their answers to the essay question. You have been given the answers written by five of the students. The class that these five students were in was a heterogeneous one. Twenty-five points is the maximum score for the question. Please grade each paper using the model answer provided. The grade is to reflect completeness and accuracy of the answer—not quality of English expression, spelling, or grammar.

Suppose that *you* grade these two answers in accordance with the instructions given above before you read any further. Record the scores that you would give the answers.

---

Now look at Table 3.1, which shows the scores actually given to all five answers, including these two. Every one of the answers receives scores spreading over about 20 points of the possible range of 25. Any one of the papers might have gotten a score as high as 18; any one might have gotten a score as low as 5. The responses of students A and B were judged to be outstandingly good by some raters, poor by others. The inconsistency of the judgments is demonstrated most forcefully. A single rating of any one of these papers

Table 3.1. Grades Given to Five Answers to Essay Question

Score	Student A	Student B	Student C	Student D	Student E
25	6	5		6	
24	2	2		4	
23	4	3		4	
22	3	9	2	5	
21	8	2		4	
20	32	21	6	24	
19	6	1		3	
18	14	11	3	12	1
17	6	8	3	2	1
16	4	2	2	4	
15	23	23	18	34	4
14	4	1	3	5	
13	2	2	3	2	1
12	4	13	9	7	6
11	1	3	6	1	2
10	6	11	33	4	25
9	1	4	9	1	5
8		6	9	3	6
7	2	3	3		
6			3	3	16
5	3		11	2	50
4			1	1	7
3		1	3		15
2			1		1
1					
0			2		1

tells us very little about how that same paper will be rated by someone else. Why is this? What makes the appraisal of an essay response so undependable?

Let us admit to start with that the dice were somewhat loaded against the graders in this little experiment. Most of them were not social studies teachers, though the majority had had some teaching experience. (Previous experience has indicated that social studies teachers will show about as much variation.) Furthermore, they had not taught the class, and did not know anything about the general level of performance in this and similar groups.

One major reason for the wide range of scores found in Table 3.1 is that different raters maintained very different standards for rating all the papers. Different raters used quite different parts of the scale of scores. Though it was most common for a rater to spread his

scores between about 5 and 20, a few awarded no grade higher than 10 to any of the answers while others assigned no grades below 15. These last two groups were operating in entirely different score ranges and showed no overlap. The best for one group was lower than the poorest for the other. Judges differed not only in the average level at which they rated the papers, but also in how much they spread out their scores. Some were very "conservative," bunching all their ratings close together, while others tended to spread them widely over the whole range. Such differences in grading standards are very real in actual school situations—as every student knows—and provide one main source for inconsistency in grading essay responses.

However, the judges were also not very consistent in the rank order in which they arranged the 5 papers. In Table 3.2 we have shown

Table 3.2. Rank Order Assigned to Each of Five Essay Questions

Rank	Student A	Student B	Student C	Student D	Student E
1	44	29	2	33	1
1.5	13	12	1	11	..
2	28	23	8	31	1
2.5	12	10	5	17	..
3	24	32	19	23	.
3.5	1	6	9	5	3
4	3	16	55	9	11
4.5	3	1	18	1	20
5	1	1	13	.	94

how often each paper was ranked first, how often second, and so on. (Tie ranks have been indicated as 1.5, 2.5, etc.) In this table we see that every one of the 5 answers was ranked first by somebody, and every answer was either last or tied for last. There is some consensus that student E wrote the poorest answer and student C the next poorest, but practically no agreement as to the relative standing of the other three. Students A, B or D could easily have been judged best of the group or only average. Thus, there is not only a marked difference in *absolute* standard from judge to judge, but also inconsistency in the *relative* judgment of one paper in comparison with the others.

Inconsistency in relative judgment is characteristic not only of different raters but also of the same rater at different times. Thus, when the evaluation class was asked to grade the papers a second time 3 weeks later (without advance notice that this was to be done), a

third of the ratings differed from the original rating by 5 points or more (out of the possible range of 25 points). Only a third of the papers kept the same rank in the group of 5 on the second grading.

The results that we have presented illustrate the situation that commonly prevails in evaluating essay responses. The responses vary in many ways and by infinitely small degrees. Raters approach them with differing standards of severity and looking for different things. As a result the evaluation of these responses is generally highly subjective and quite unreliable. We shall consider later in the chapter what can be done to deal with these very real problems.

### THE OBJECTIVE TEST

The objective test includes a variety of forms of test tasks having in common the characteristic that the correct answer, usually only one, is determined when the test item is written. The word "objective" in objective test refers only to the scoring of the answers; the choice of content and coverage of an objective test is probably as subjective as the choice of content and coverage of an essay test, and for some types of items there is subjective judgment involved in the original decision as to what is the correct answer. Common forms of objective test items are shown below.

#### True—False

T F

The Constitution states that United States Senators shall be elected for terms of 4 years.

#### Multiple Choice

A law passed by a legislature to punish a person without a court trial would be called

- A. an ex post facto law.
- B. a bill of attainder.
- C. a writ of habeas corpus.
- D. a warrant.

#### Completion

The right to vote is called (suffrage).

#### Matching

Column I—Documents

- B The Mayflower Compact
- C The Petition of Right
- A The Magna Charta

Column II—Dates

- A. 1215
- B. 1620
- C. 1628
- D. 1689
- E. 1776

The essential features of a test made of objective items, as distinct from an essay test, are that the examinee

1. Operates within an almost completely structured task.
2. Selects one of a limited number of alternatives.
3. Responds to each of a large sample of items.
4. Receives a score for each answer according to a predetermined key.

Again, let us examine these characteristics to see the advantages and disadvantages of each. In large measure, they are the reverse of those discussed for essay examinations.

*The Task Is Completely Structured.* The examinee does not have a chance to organize and define the problem for himself. On the debit side, this means that a test of this sort is not useful for appraising skills of organizing and structuring ideas. On the credit side, we are more sure that each examinee is presented with the same problem. "Discuss the Articles of Confederation," can carry quite different meanings to different pupils.

*The Examinee Selects from Among Given Alternatives.* In most types of objective item, the possible alternatives are completely specified. (This is not the case with the completion type of item, and in that respect it is on the boundary line, approaching the short free-response type of question.) Where the alternatives are all provided, the student is only required to recognize the right answer, not to produce it by his own efforts. This has been criticized as representing a lower level of intellectual process, and one that is less true to life. How valid this criticism is probably depends upon how skillfully the objective items are written, and how much they manage to get away from the words of the text and simple memory of factual materials. When an objective test item presents a new problem that must be solved by recalling and applying facts or principles previously learned, this type of item can require just as active recall as any essay question.

Another outcome of the limited set of answer choices is that an examinee can be expected to get some answers right by guessing. This becomes a problem particularly for true-false questions in which there are only two choices. Tossing a coin would give 50 per cent right on the average, and people would get different scores to some extent because they were lucky or unlucky coin tossers. The problem of guessing is serious in a short test with few answer choices, but chance successes tend to even up in the long run if there are enough items, if enough time is given for everyone to complete the test, and if instructions about guessing can be made sufficiently definite so that all examinees will adopt the same policy.



*The Sample of Items Is Large.* Since each item is brief, many items can be included. These can be spread more evenly over the topics to be covered and a more representative sampling can be obtained. This reduces the role of luck, of the individual just happening to have reviewed a particular topic. As a consequence of the inclusion of many separate items, the score from a well-made objective test is likely to be more accurate than that from an essay test, so that two separate tests of an individual based on the same content areas will rank him in more nearly the same place in his group.

*Each Item Has a Predetermined Key.* The key is established once and for all by the test maker at the time the test items are written. This means that scoring the test is a routine clerical task and can be done by a person who knows nothing about the subject matter of the test or even by one of the electrical test-scoring machines on the market. The saving in time to score the test is very substantial, but it must be remembered that much of that saving will have been used up in preparing the test. Writing clear and unambiguous objective test items is a fairly demanding literary task.

The economy in time is less important than the uniformity in evaluating answers that results. The score will be the same whoever scores the test, once the key has been agreed upon. The score will be the same no matter who it was that chose the answers. Teacher's pet or hellion, Spencian specialist or scribbler, if they choose the same answer they get the same score.

#### SUMMARY COMPARISON

The issues we have been discussing are summarized in tabular form below. In each case a plus sign is placed in the column of the test pattern that would be judged superior with respect to that factor.

Factor	Essay	Objective
Provides opportunity to test student's ability to select, organize, and integrate *	+	
Requires student to produce answer and not just recognize it	+	
Is free from factors of skill in expression and penmanship		+
Is free from opportunities for bluffing		+
Is free from opportunities for guessing	+	
Provides an adequately representative sample of the topics covered		+
Can be prepared quickly	+	
Can be scored quickly		+
Can be scored routinely by a clerk		+
Can be scored with high consistency from scorer to scorer		+

The balance of importance between these factors will vary from situation to situation. It is clear that neither type has exclusive claim to all the advantages. In evaluating the work of his class, the teacher needs to use both kinds of testing procedures.

## EFFECTIVE USE OF THE ESSAY EXAMINATION

Because of their advantages in evaluating abilities to organize an answer to a question, recall and select relevant information, and present it logically and effectively, essay examinations should continue to be used in the evaluation of student performance. If they are to be used, the teacher should have some guiding principles as to when to use them and what he can do to overcome their common weaknesses. These weaknesses are found partly in the format of the questions and partly in the process of evaluating the answers produced by the students.

### WHEN TO USE ESSAY EXAMINATIONS

The factors that make it appropriate to use an essay examination are in part very immediate practical ones, in part more fundamental theoretical considerations.

*Immediate Practical Considerations.* The most obvious practical reason for using an essay examination is to save time. It takes a number of hours to prepare a good objective test. When the class group is small, there will be few papers to read and an essay examination may actually save time. Moreover, when time to prepare an examination is limited the teacher can substitute reading time *after* the examination for preparation time *before* the examination. Since many fewer essay questions than objective items are required for a given amount of testing time, the teacher may find it easier to construct a good essay test than a good objective test. However, it should be emphasized that making good tests of any kind requires considerable thought and effort on the part of the person writing the questions. A teacher cannot expect to produce good tests of any kind if he dashes off the questions a half-hour before the test is to be given.

A consideration that may be compelling in some cases is lack of reproduction facilities for running off copies of the test. Then a set of essay questions written on the blackboard is a practical solution. In such a situation it would probably be wise to use some short free-answer questions requiring only a few words or sentences for an answer as well as those in true essay form requiring extended answers. This will permit a wider and more adequate sample of the students' achievement. Another practical solution is to read objective questions

to the class. This procedure will work for rather alert students and for fairly simple items but it tends to be inefficient in requiring repetitions of the items and it requires a somewhat special ability to remember the total item well enough to indicate an answer. With more complex items, the teacher will find that reading the items to the class becomes less satisfactory.

A third point that is sometimes made is that essay questions are less demanding upon the skill of the teacher. It is probably true that ambiguities and poor expression are more apparent in an objective item, but confusion as to what is wanted in response to an essay question can also be substantial. In many educational settings, the student must know the person who wrote the essay question in order to write an acceptable answer. Many of the faults in writing objective items can be avoided once they have been pointed out, so that it seems more desirable to improve item-writing skills than to resort to essay questions as a defense.

*More Basic Theoretical Issues.* The functions that can be appraised better by an essay question than by short-answer or objective questions are abilities to select, relate, and organize, to create essentially new patterns and to use language to express one's ideas. For example, objective 5 on our blueprint on p. 33, "can express generalizations and concepts in his own words," can be measured only by an essay item since an objective item does not permit the student to use his own words. There would be little justification for using essay items to evaluate objectives one through three on our blueprint since these objectives require the reproduction of factual information. The essay question is an inefficient way to measure factual information that could be more effectively and efficiently measured by a series of objective items.

Merely phrasing a question in the essay form does not automatically insure that the abilities to select and organize, to create new syntheses, to make new applications, or the other so-called higher mental abilities will be assessed. Most of the essay tests given in elementary and secondary schools and colleges measure nothing more than the ability to reproduce facts. In order to assess the abilities that are best measured by essay questions, the questions must be carefully phrased to require an application or creative synthesis of what has been taught. Thus, question A tests only information.

#### *Question A*

What rights are guaranteed to the people under the first amendment to the Constitution?

### Question B

A newspaper, *The Evening Standard*, published a series of articles on the city government of Townsville. In one article, the reporter for the paper stated that the mayor of Townsville was incompetent and inefficient and did not spend enough time in his office to take care of city affairs. The mayor sued *The Evening Standard* in court for libel stating that the article made him look bad to the people of the city and reduced his effectiveness as mayor. What decision could be expected from the courts? Why?

Question B, by contrast, requires identification and selection of the proper items of information, and their application to the solution of a new problem. Question B seems more clearly appropriate for an essay examination.

In the early days of objective testing, some studies were carried out that showed that the prospect of an essay examination leads to study activities emphasizing the interrelationships of facts and principles in an area whereas the prospect of an objective examination leads to the memorizing of discrete details. There is little recent evidence on this point, and we wonder whether the relationship was a *necessary* one or merely a reflection of the low quality of the objective tests to which the groups had been exposed. This finding certainly points out a *potential* weakness of objective tests, and one escape from this weakness is to use essay tests. We suspect that study habits depend less upon the form of the test exercises than upon the type of objective that is emphasized—whether the items are objective or essay.

*Variants on the Essay Examination.* Values claimed for the essay examination are those of appraising ability to organize materials and to use language effectively to express the resulting organization. However, in the usual scheduled essay examinations these functions may become submerged because (1) differences in knowledge of the basic facts hide differences in ability to organize those facts and (2) time pressures hide the quality of the individuals' written expression.

Two variations may be considered that appear likely to bring out the factors in which we are particularly interested. One is to give an "open book" examination, in which every individual has access to any basic data present in his text, his notes, or other sources. Memory of facts is then reduced as a factor entering into individual performance, and ability to locate, select, and use the facts is brought to the fore.

The second variation is to give the problems as an out-of-class examination with unlimited time. This minimizes time pressure, and

makes the test more nearly a pure power test—power both with respect to organizing ability and with respect to written expression. We do, of course, introduce a new problem, since we are less able to guarantee the integrity of the written material turned in. When the examination is used against rather than for the pupil, illicit help is likely to become a serious problem.

### IMPROVING ESSAY TESTS

We have already pointed out in a previous section that essay tests can be improved by limiting their use to those objectives that are best measured by the essay format. There is not much that a teacher can do to overcome the weakness of essay tests that arises from the limited number of essay questions that can be presented to students in a given period of time except to give several essay tests during the school semester or year. A teacher can do much to overcome some of the other weaknesses of the essay test by (1) writing good essay questions, and (2) improving his methods of evaluating the answers. In the next section we will give some guides to writing better essay questions. Following that is a section suggesting ways to improve the scoring of answers to essay questions.

### IMPROVING THE CONTENT OF AN ESSAY TEST

The following paragraphs present and discuss several suggestions for improving the questions that go into an essay test. These are not scientifically established principles, but they reflect the judgment of experienced test makers.

1. *Before starting to Write the Essay Question, Have in Mind Explicitly What Mental Processes of the Student You Want to Bring Out by the Question.* If you want to use the essay question to determine the extent to which a student can use his information, then the question must be phrased in such a way that the student must do such things as solve a problem that has not been directly taught, or point out relationships that have not been explicitly pointed out before.

2. *In General, Start Essay Questions with Such Phrases as "Compare," "Contrast," "Give the reasons for," "Present the arguments for and against," "Give original examples of," and "Explain how or why."* These words will help to present tasks requiring the student to select, organize, and apply his knowledge. Don't start essay questions with such words as "what," "who," "when," and "list." These words are likely to present tasks requiring only the reproduction of information.

3. *Write the Essay Question in Such a Way That the Task Is Clearly and Unambiguously Defined for Each Examinee.* A question such as "Discuss the factors and influences that led to the writing and adoption of our Constitution," is global, vague, and ambiguous. First, what does the teacher mean by the word "discuss"? Second, does the teacher want the student to start with the Magna Charta in 1215 or with the settlement of the colonies or with the end of the Revolutionary War? Third, does the teacher want the student to stop with the beginning of the Constitutional Convention in 1787 or with the ratification of the Constitution? Fourth, what does the teacher mean by "factors and influences?" The score that the student receives for his answer is likely to depend to a large extent on how lucky he is guessing what the teacher wanted.

A better way to phrase this question so that each examinee will interpret the question in the same way would be:

Explain how each of the following influenced the provisions written into our Constitution by the delegates to the Constitutional Convention.

- A. The Magna Charta, the Petition of Right, and the English Bill of Rights.
- B. The fear of tyranny or rule by one man or one group.
- C. The problems that arose in trying to operate under the provisions of the Article of Confederation.
- D. The fear of the small states that they would be controlled by the large states.
- E. Business rivalries between states.

The question as it has been rephrased guarantees a more common basis for response. In one sense it breaks the one question up into five. The analysis also makes clear that on the original question (and also the revised one) students will require a relatively long time to write an adequate answer.

4. *The Words "What do you think," "In your opinion," or "Write all you know about . . ." Almost Never Belong in an Essay Question to Measure Academic Achievement.* The use of these phrases is common on teacher-made essay tests. But when a teacher asks: "Why do you think that the Articles of Confederation provided a poor basis for the formation of our central government?" he is not really interested in the student's opinion. He actually wants to determine whether the student knows the fundamental weaknesses of the Articles of Confederation, as stated by the teacher or text. Therefore the question would be better if written: "Why did the Articles of Confederation prove to be unworkable as a framework for our national government?"

The only time when the use of "you," "in your opinion," or "do you think" is justified in an essay question (or any other type of test question) is when the purpose of the question is to obtain an expression of attitudes (which really cannot be graded) or to determine how good a logical defense a student can make of the position that he has taken. In the latter instance, the teacher should *not* be particularly interested in which position the student takes and should evaluate the answer given only on the basis of how well the student defends or supports his position.

5. *Be Sure That the Students Do Not Have Too Many or Too Lengthy Questions to Answer in the Time Available.* An essay test should not be a test of speed of writing. Good essay questions demand that the student consider the question, think about his answer, then write it. These processes take time and the younger the student or the more complex the question, the longer is the required time. In order to answer adequately the revised question on p. 54, the typical eighth grader would probably need from 45 to 60 minutes. In most essay tests given in the classroom, three to five such questions are given to be answered in a single classroom period. This practice may encourage both sloppy thinking and sloppy writing on the part of the student.

6. *Do Not Use Both Essay and Objective Questions in the Same Test when the Time for Testing is Limited.* Quite frequently teachers use both objective and essay questions on the same test. It is not unusual to see a teacher-made test consisting of thirty to fifty multiple-choice questions and one to three essay questions, all of which are to be answered in a 50-minute period. This practice is undesirable first because there is not enough time for the student to answer adequately all of the questions and second because there are very difficult problems in combining the scores on the two different kinds of items. (See Chapter 17.)

7. *Have Each Examinee Answer the Same Questions. Don't Offer a Choice of Questions to be Answered.* When an essay examination is being used to appraise achievement of the objectives of a common program of study, each examinee should be required to answer the same questions. Giving a choice of questions reduces the common base upon which different individuals may be compared. It adds one further source of variability to the subjectivity and inaccuracy that already exist. A choice of questions may have a public-relations value with the examinees, but it has no justification from the point of view of effective measurement.

## SCORING ESSAY EXAMINATIONS

A number of steps may be taken to mitigate the subjectivity and reduce some of the biases in evaluating the answers to an essay examination. These are mostly attempts to break up the process of evaluation into a series of more specific, fractionated judgments made upon a common base and applied to an anonymous product. Specific suggestions are outlined below.

1. *Decide in Advance What Factors Are to Be Measured. If More than One Distinct Quality Is to Be Appraised, Make Separate Evaluations of Each.* If facts are considered important, score for facts. If organization is important, give a rating upon organization. If mechanics of English, sentence structure, spelling, punctuation, etc., are considered a significant outcome, give a rating upon mechanics. However, do not contaminate the rating for knowledge or understanding with appraisal of mechanics. It is hard to isolate quality of organization from extent of factual information, but if the essay question is to serve its distinctive purpose an attempt should be made to do so.

2. *Prepare a Model Answer in Advance, Showing What Points Should Be Covered and How Many Credits Are to Be Allowed for Each.* This will provide a common frame of reference for evaluating each paper. After the preliminary model has been prepared, it should be checked against a sample of student responses to the question. The model and the scoring scheme should be modified in the light of these answers. They can now be used as the yardstick for assigning credits to each paper in turn.

3. *Read All Answers to One Question before Going on to the Next.* A more uniform standard can be maintained for a single question and for a short period of time. There is more chance to compare one person's answer with another's and thus to build up a "feel" for the answers. There is less contamination of judgment by what that same examinee had written on the previous question.

4. *Grade the Papers as Nearly Anonymously as Possible.* The less you know about *who* wrote an answer, the more objectively you can grade *what* was written.

5. *Greater Reliability Can Be Obtained by Averaging Independent Ratings.* If the importance of the test merits the expenditure of the extra effort, a more dependable appraisal can be obtained by having one or more additional raters each give an independent rating of the responses.



## SUMMARY STATEMENT

Evaluation of pupil achievement is one of the teacher's important responsibilities. In view of the many functions that tests serve in motivating and directing learning, and in view of the disservice that may be done the pupil from poorly conceived or executed evaluation instruments, it is important that the teacher's evaluation devices be well thought out and well made. Both written tests and a variety of informal appraisals are needed to evaluate completely the objectives of the modern curriculum.

For any type of written test, it is desirable to have a definite plan in advance of preparing the test items. The development of such a plan requires an analysis of the outcomes one is trying to achieve in the teaching of a particular course or unit and of the significant segments of content through which those objectives are to be realized. A statement of objectives useful for guiding the construction of test items must be phrased in terms of pupil behaviors—specific things that the pupil is supposed to be able to do—rather than in broad generalizations. In addition, the plan should include the allocation of test items among the content areas and objectives, the types of items to be used, the total number of items in the test, and specifications for the spread of item difficulties.

Both essay and objective tests should be used to evaluate pupil achievement. The essay test is easier to prepare and has certain advantages in appraising ability to recall information, select relevant material, and organize it into an integrated answer. However, the objective test has marked advantages in freedom from such irrelevant factors as quality of handwriting or of English usage, in breadth of sampling of the desired outcomes of teaching, and in ease and objectivity of scoring.

Essay questions can be improved by phrasing the question so as to present a well-defined task to the student and by providing conditions for scoring that reduce as far as possible the subjectivity of grading.

## REFERENCES

1. Bloom, Benjamin S., Max D. Engelhart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl, *Taxonomy of educational objectives: the classification of educational goals: handbook 1, cognitive domain*, New York and London, Longmans, Green, 1956.

## SUGGESTED ADDITIONAL READING

- Bloom, Benjamin S., Editor, *Taxonomy of educational objectives, Handbook I, Cognitive domain*, New York, Longmans, Green, 1956.
- Dressel, Paul L., and Lewis B. Mayhew, *General education: explorations in evaluation*, Washington, D. C., American Council on Education, 1954, Chapters 3-8.
- French, Will, *Behavioral goals of general education in high school*, New York, Russell Sage Foundation, 1957.
- Harris, Chester W., Editor, *Encyclopedia of educational research*, 3rd ed., New York, Macmillan, 1960, pp. 650-657, 1506-1514.
- Kearney, Nolan C., *Elementary school objectives*, New York, Russell Sage Foundation, 1953.
- Lindquist, E. F., Preliminary considerations in objective test construction, Chapter 5 in E. F. Lindquist, Editor, *Educational measurement*, Washington, D. C., American Council on Education, 1951.
- Odell, C. W., *How to improve classroom testing*, rev. ed., Dubuque, Iowa, William C. Brown, 1958, Chapters III, IV, V, and VI.
- Smith, Eugene R., et al., *Appraising and recording student progress*, New York, Harper, 1942, Chapters 1 and 2.
- Stalnaker, John M., The essay type of examination, Chapter 13 in E. F. Lindquist, Editor, *Educational measurement*, Washington, D. C., American Council on Education, 1951.
- Thomas, R. Murray, *Judging student progress*, 2nd ed., New York, Longmans, Green, 1960, Chapter 2.
- Vaughn, K. W., Planning the objective test, Chapter 6 in E. F. Lindquist, Editor, *Educational measurement*, Washington, D. C., American Council on Education, 1951.

## QUESTIONS FOR DISCUSSION

1. Prepare a statement of the objectives for a course, or a unit within a course, that you are teaching or plan to teach.
2. Which of the objectives in 1 could be measured effectively by a written test? Which only partially or not at all. Why is a written test inadequate for these? How might these objectives best be appraised?
3. Based on the objectives identified in the first part of question 2 and a course outline, prepare a blueprint for a test to evaluate the unit or course.
4. In a junior high school, one teacher takes complete responsibility for preparing the common final examination for all the classes in general science. He makes the examination up without consulting the other teachers. What advantages and disadvantages do you see in this procedure?
5. It has been said that one of the goals of the music program in an elementary school is to "increase the sensitivity of pupils to music in its different forms." How could this goal be defined so that progress toward it could be measured?

6. Students are sometimes heard to remark: "You can't get a good mark on Miss X's tests unless you really know Miss X." What does this remark imply about Miss X's tests?

7. On p. 49 is a list of factors that have been presented as favoring either essay or objective tests. Do you agree with the classification given there? Which are the most important factors? What other points should be considered in deciding which type of test to use for the final examination in a particular course?

8. Criticize the following features of an essay test planned for a ninth-grade social studies class:

- a. There will be 10 questions on the test.
- b. Each student will answer any 5.
- c. Each question will have a value of 20 points.
- d. One point will be taken off for each misspelled word and each grammatical error.
- e. A 5-point bonus will be given for neatness.
- f. Time for the test will be 40 minutes.

9. Criticize and revise each of the following essay questions:

- a. Discuss the increase in juvenile delinquency since World War II.
- b. Discuss government support of farm prices.
- c. Discuss the "cold war."

10. For what types of objectives would an open-book essay examination be appropriate? What would be the advantages and disadvantages of such an examination, as compared with the usual essay examination?

## Chapter 4



# Preparing Objective Tests

### INTRODUCTION

The objective type of test item was developed in order to overcome some of the disadvantages of the essay test discussed in Chapter 3. As we pointed out in that chapter, there is still a good deal of argument about the relative merits of the two types of test. Those who object to the objective type of test say that it emphasizes factual material, encourages piecemeal memorization of unimportant details, permits too much guessing of the correct answer, ignores the higher mental processes, neglects the more important educational objectives, and never gives the student any practice in writing. Except for the last objection, we have discussed the other criticisms in Chapter 3. As for the last objection, we might well raise the question as to whether the testing period is the place to give students practice in writing and whether the kind of writing practice provided by most essay tests encourages (or discourages) good writing.

As we have stated before, the question of which kind of test to use is not an either-or question. Both essay and objective tests can be used to advantage in the classroom. A poorly constructed test of either kind can inhibit or misdirect learning. The problem then is to construct good tests. In this chapter we will consider methods of improving and using the objective type of item and of analyzing and using the results of objective tests.

### WRITING THE ITEMS FOR AN OBJECTIVE TEST

Writing good test items is an art. It is a little like writing a good sonnet and a little like baking a good cake. The operation is not quite so free and fanciful as writing the sonnet; it is not quite so standardized as baking the cake. It lies somewhere in between. So a discussion of item writing lies somewhere between the exhortation to the poet to go out and express himself and the precise recipes of a good cookbook. The point we wish to make is that there is no exact

science of test construction. The guides and maxims that we shall offer are not tested out by controlled scientific experimentation. Rather, they represent a distillation of practical experience and professional judgment. As with the recipe in the cookbook, if carefully followed they yield a good product.

We shall first present some suggestions that apply to almost any type of objective item. Then we will consider specific item types, indicating some of the general virtues and limitations of that type of item and giving more specific suggestions for writing and editing. A number of the principles that we set forth will seem very obvious. However, experience in reviewing and editing items indicates that these most obvious faults are the ones that are most frequently committed by persons who try to prepare objective tests. Thus, it hardly seems necessary to insist that a multiple-choice item must have one and only one right answer, and yet items with no right answer or several occur again and again in tests that are carelessly prepared.

#### GENERAL MAXIMS FOR ITEM WRITING

1. *Keep the Reading Difficulty of Test Items Low* in relation to the group who are to take the test, unless the purpose is to measure verbal and reading abilities. Ordinarily you do not want language difficulties to interfere with a pupil's opportunity to show what he knows.

#### Example

*Poor:* What was the ostensible reason for requesting the states to designate one or more of their constituents as representatives to attend a general convention to meet in Philadelphia in 1787?

- A. To draft a new Constitution.
- B. To raise money to pay off Revolutionary War debts.
- C. To settle commercial disputes among the states.
- D. To revise the Articles of Confederation.

*Better:* When the states were asked to send representatives to a general convention to meet in Philadelphia in 1787, they were told that these representatives would be asked to

- A. draft a new Constitution.
- B. raise money to pay off Revolutionary War debts.
- C. settle commercial disputes among the states.
- D. revise the Articles of Confederation.

2. *Do Not Lift a Statement Verbatim from the Textbook.* This places a premium upon rote memory with a minimum of understanding. Also the statement may have little or no meaning when it is

removed from the context. A statement can at least be paraphrased. Better still, in many cases it may be possible to imbed the specific knowledge in an application.

### Example

*Poor:* T F The House of Representatives shall be composed of members chosen every second year.

*Better:* T F A United States Representative elected for a full term of office to begin in 1961 would end his term in 1963.

3. *If an Item Is Based on Opinion or Authority, Indicate Whose Opinion or What Authority.* Ordinarily statements of a controversial nature do not make good items, but there are instances where knowing what some particular person thinks may be important for its own sake. The student should presumably be acquainted with the viewpoint of his textbook or instructor, but he should not be placed in the position of having to endorse it as indisputable fact.

### Example

*Poor:* T F The Declaration of Independence influenced later political developments more than any other document.

*Better:* T F According to your textbook, the Declaration of Independence influenced later political developments more than any other document.

4. *In Planning a Set of Items for a Test, Take Care That One Item Does Not Provide Cues to the Answer of Another Item or Items.* The second item below gives cues to the first.

### Example

1. Under the provisions of the Constitution, the judicial branch of our National Government is given the power to
  - A. enforce the laws.
  - B. interpret the laws.
  - C. make the laws.
  - D. repeal the laws.
2. The interpretation of laws by the judicial branch of our National Government has been one method used to
  - A. keep the powers of government in the hands of the people.
  - B. prevent a weak president from being dominated by a strong Congress.
  - C. guarantee Constitutional rights to all citizens.
  - D. keep the Constitution flexible enough to meet changing social, political, and economic conditions.

5. *Avoid the Use of Interlocking or Interdependent Items.* The answer to one item should not be required as a condition for solving the next item. This is the other side of the principle stated in 4 above. Every individual should have a fair chance at each item as it comes. Thus, in the example shown below, the person who does not know the answer to the first question is in a very weak position as far as attacking the second one is concerned.

*Example*

1. The name of the first written constitution in the American colonies was the (Fundamental Orders of Connecticut).
2. This constitution was drafted in the year (1639).

6. *In a Set of Items, Let the Occurrence of Correct Responses Follow Essentially a Random Pattern.* Avoid favoring certain responses, i.e., either true or false, or certain locations in a set of responses. Do not have the responses follow any systematic pattern.

7. *Avoid Trick and Catch Questions,* except in the rare case in which the test has a specific purpose of measuring ability to keep out of traps. Trick questions are likely to mislead the abler or better-informed student, who knows enough to be caught by the trap. If they do this, they defeat the basic purpose of the test, which is to identify levels of knowledge and understanding.

*Example 1*

*Poor:* T F The term of office for all senators is 6 years.

(The item is keyed true but the student who knows the most about government is likely to get it wrong because a senator who is elected or appointed to take the place of a senator who dies serves only the unexpired time. Also, among the first group of senators at the time the Constitution was adopted, some served only 2 years, some 4, and some 6.)

*Better:* T F The Constitution states that the term of office for senators shall be six years.

*Example 2*

*Poor:* T F On May 25, 1787, fifty-five delegates from twelve states met to revise the Articles of Confederation.

(This was keyed false because all fifty-five were not present on May 25. No revision is shown for this item because the idea being tested is considered so insignificant that it would be better not to use the item.)

8. *Try to Avoid Ambiguity of Statement and Meaning.* This is a general admonition, somewhat like "Sin no more," and it may be no

more effective. However, it is certainly true that ambiguity of statement and meaning is the most pervasive fault in objective test items. Many of the specific points already covered and many of those still to be covered deal with specific aspects of the reduction of ambiguity.

### Example

*Poor:* In the Constitution, the composition of Congress was established in order to

- A. maintain balance of power between the large and small states.
- B. protect the interests of propertied classes.
- C. get the delegates to accept and sign the Constitution.
- D. provide for a stronger central government.

The keyed answer to the above question was A, but the examinee trying to answer the item is faced with several problems. First of all, what does the writer of the item mean by "the composition of Congress?" Does he mean the division of Congress into two houses, the basis for determining representation in Congress or the qualifications of the members of Congress? Does the writer of the item want the student to give the immediate reason for the compromise or the ultimate reason? Actually the writer of this item was trying to determine whether the student knew why the Constitution provided for a Congress made up of a House of Representatives with proportional representation from each state and a senate with equal representation from each state.

But even if the student guesses correctly what the item writer had in mind when he wrote the item, he is likely to have difficulty with the answer choices.

A case can be made for each of the answer choices being correct. All of the compromises at the Constitutional Convention had two aims: to provide for a stronger central government and, at the same time, to draft a document that the states would be willing to accept. There is some truth in choice B because the large states feared that the small states would pass laws interfering with business and property. Of all the answer choices, the keyed answer A is probably the least correct since the purpose was not to maintain exact balances of powers between large and small states but to grant some concessions to each.

The item needs to be sharpened up in several respects. The example below would appear to test the same knowledge and to provide less occasion for misunderstanding of what the examiner was trying to say.

*Better:* At the Constitutional Convention, the delegates agreed to give each state equal representation in the Senate and proportional representation in the House of Representatives in order to

- A. satisfy the conflicting demands of the large and small states.
- B. protect the rights of the sovereign states.
- C. make the legislative branch of the central government the strongest.
- D. keep the government in the hands of all the people.



9. *Beware of Items Dealing with Trivia.* An item on a test should appraise some important item of knowledge or some significant understanding. Avoid the type of item that could quite justifiably be answered, "Who cares?" Ask yourself in each case whether knowing or not knowing the answer would make a significant difference in the individual's competence in the area being appraised.

### Example

*Poor:* A census every 10 years was provided for in the Constitution in Article I Section

- A. 1
- B. 2
- C. 3
- D. 4

*Better:* The reason the framers of the Constitution provided that a national census should be taken every 10 years was to

- A. obtain information needed by Congress to carry out its duties.
- B. determine how many Representatives each state should have.
- C. determine how rapidly the country was growing.
- D. obtain accurate information for use by government and industrial agencies.

### TRUE-FALSE ITEMS

The true-false item has had a popularity in teacher-made objective tests far beyond that warranted by its essential nature. This has probably happened because bad true-false items can be written quickly and easily. To write good ones is quite a different matter.

Even when they are well written, true-false items are relatively restricted in the types of educational objective they can measure. They should be limited to statements that are unequivocally true or demonstrably false. For this reason, they are adapted to measuring relatively specific, isolated, and often trivial facts. They can also be used fairly well to test meanings and definitions of terms. But items testing genuine understandings, inferences, and applications are usually very hard to cast in true-false form. The true-false item is particularly open to attack as fostering piecemeal, fractionated, superficial learning and is probably responsible for many of the attacks upon the objective test. It is also in this form of test that the problem of guessing becomes most acute.

The commonest variety of true-false item presents a simple declarative statement, and requires of the examinee only that he indicate whether it is true or false.

*Example*

T F The Articles of Confederation provided for a strong central government.

Several variations have been introduced in an attempt to improve the item type. One simple variation is to underline a part of the statement, viz., "strong" in the above example. The instructions indicate that this is the key part of the statement and that it determines whether the statement is true or false. That is, the correctness or appropriateness of the rest of the statement is guaranteed. The examinee can focus his attention upon the more specific issue of whether the underlined part is compatible with the rest of the statement. This seems to reduce guessing and make for more consistent measurement.

A further variation is to require the examinee to correct the item if it is false. This works well if combined with the underlining described above but is likely to be confusing if no constraints are introduced in the situation. Our example could be corrected by changing "Articles of Confederation" to "Constitution," by changing "strong" to "weak," or by changing "central" to "state." Requiring that the item be corrected reduces guessing and provides some further cue to the individual's knowledge.

Generally, the true-false type of item tends to be most useful when it is based on some given stimulus material such as a chart, map, graph, table, or reading passage and when the student responds to the item only in terms of the given material. This type of true-false item has been used effectively in testing ability to interpret data of different kinds. However, in this case, the format is generally changed by requiring the student to answer in four or five categories such as definitely true, probably true, insufficient data to determine whether it is true or false, probably false, and definitely false. In this format the item is more like a multiple-choice item than a true-false item.

## CAUTIONS IN WRITING TRUE-FALSE ITEMS

1. *Be sure that the Item as Written Can Be Unequivocally Classified as Either True or False.* One of the most common weaknesses in true-false items is that the person who knows the most about the content may find it difficult to judge whether the item is true or false. This is particularly likely to happen with items that were intended to be true statements. The student who knows the most about the content can often think of a number of exceptions or reasons why the statement is not universally true. Consider the following example.

*Example*

*Poor:* T F The presidential candidate who receives the majority of votes is elected President.

The item was keyed true but strictly speaking it is not true. The candidate must receive the majority of *electoral* votes but not necessarily the majority of the *popular* vote. It is the higher-achieving student who is likely to know about both the electoral votes and the popular vote and he is likely to mark the item false because it does not specify electoral votes. The item would be better if it were revised as follows:

*Example*

*Better:* T F The presidential candidate receiving a majority of the electoral votes is elected President.

2. *Beware of "Specific Determiners,"* words that give cues to the probable answer, such as all, never, usually, etc. Statements that contain "all," "always," "no," "never," and such all-inclusive terms represent such broad generalizations that they are likely to be false. Qualified statements involving such terms as "usually" or "sometimes" are likely to be true. The test-wise student knows this, and will use these cues, if he is given a chance, to get credit for knowledge he does not possess. "All" or "no" may sometimes be used to advantage in *true* statements, because in this case using the determiner as a cue will lead the examinee astray.

*Example*

*Poor:* T F All sessions of Congress are called by the President.

*Better:* T F All persons elected to the House of Representatives must be at least 25 years old.

3. *Beware of Ambiguous and Indefinite Terms of Degree or Amount.* Expressions such as "frequently," "greatly," "to a considerable degree," and "in most cases" are not interpreted in the same way by everyone who reads them. Ask a class or other group what they think of when you say that something happens "frequently." Is it once a week or once an hour? Is it 90 per cent of the time or 50 per cent? The variation will be very great. (Ed.: How great is very great?) An item in which the answer depends on the interpretation of such terms as these is an *unsatisfactory one*.

*Example*

*Poor:* T F The Supreme Court is frequently required to rule on the constitutionality of a law.

*Better:* T F The Supreme Court has the power to declare a law unconstitutional.

4. *Beware of Negative Statements and Particularly of Double Negatives.* The negative is likely to be overlooked in hurried reading of an item, and the double negative is hard to read and confusing.

#### Example

*Poor:* T F The Constitution does not provide that no state law can deny a citizen the right to vote.

*Better:* T F The Constitution grants to each state the right to make laws specifying the qualifications for voting in that state.

5. *Beware of Items that Include More than One Idea in the Statement, Especially If One Is True and the Other Is False.* This type of item borders on the category of trick items. It places a premium on care and alertness in reading. The reader must not restrict his attention to one idea to the exclusion of the other or he will be misled. The item tends to be a measure of reading skills rather than knowledge or understanding of subject content.

#### Examples

*Poor:* T F The President has the power to make treaties with foreign countries, but the Senate must approve them by a majority of votes.

*Better:* T F The Senate must approve a treaty with a foreign country by a majority of votes.

*Poor:* T F No person shall be elected to the office of president more than twice, but a person who has acted as president for 2 years or more shall be eligible for re-election for at least two full terms.

*Better:* T F A person who has acted as president for 2 or more years can be re-elected twice.

(In each of the poor items, the first statement is true and the second one is false.)

6. *Beware of Items Where the Correct Answer Depends upon One Insignificant Word, Phrase, or Letter.* Each test item should measure an important aspect of the student's achievement; therefore each true-false item should require the student to react to important ideas and should not require him to be a proofreader. Many teachers try to obtain a spread of scores on a test by introducing items that require the student to examine each word and each letter in the word in order to arrive at the correct answer. For example, the item, "Ulysses Sampson Grant was President of the United States from 1869 to 1877," appeared as a true-false item on a sixth-grade social studies test and was

keyed false because Grant's middle name was Simpson, not Sampson. Surely, knowing Grant's middle name is not a significant aspect of achievement in sixth-grade social studies; however, if it is, then the item should be written so that attention is drawn to the middle name of Grant; e.g., "T F Ulysses Grant's middle name was Sampson."

7. *Beware of Giving Cues to the Correct Answer by the Length of the Item.* There is a general tendency for true statements to be longer than false ones. This is a result of the necessity of including qualifications and limitations to make the statement true. The item writer must be aware of this trend and make a conscious effort to overcome it.

#### SHORT-ANSWER AND COMPLETION ITEMS

The short-answer and the completion item tend to be very nearly the same thing, differing only in the form in which the problem is presented. If it is presented as a question it is a short-answer item, whereas if it is presented as an incomplete statement it is a completion item.

#### Example

*Short Answer:* In what colony was the first representative assembly in America established?

*Completion:* The first representative assembly in America was established in the colony of (Virginia).

Items of this type are well suited to testing knowledge of vocabulary, names or dates, identification of concepts, and ability to solve algebraic or numerical problems. Numerical problems that yield a specific numerical solution are "short answer" in their very nature. The measurement of more complex understandings and applications is difficult to accomplish with items of this type. Furthermore, evaluation of the varied responses that are given is likely to call for some skill and to introduce some subjectivity into the scoring procedure.

#### MAXIMS CONCERNING COMPLETION ITEMS

1. *Beware of Indefinite or "Open" Completion Items.* In the first example, on p. 70, there are many words or phrases that give factually correct and reasonably sensible completions to the statement, i.e., "arrested," "imprisoned," "acquitted," "critical of the government," "from New York," "a publisher." The problem needs to be more fully defined, as is done in the revised statement.

*Example*

*Poor:* The man whose case won freedom of the press for our country was (Zenger).

*Better:* The name of the man whose case won freedom of the press for our country was (Zenger).

2. *Omit Only Key Words.* Do not leave the verb out of a completion statement unless the purpose of the item is to measure knowledge of verb forms. The blank in a completion item should require the student to supply an important fact.

*Example*

*Poor:* The Constitutional Convention (met) in Philadelphia in 1787.

*Better:* The Constitutional Convention met in Philadelphia in the year (1787).

3. *Don't Leave Too Many Blanks in a Statement.* Overmutilation of a statement reduces the task of the examinee to a guessing game or an intelligence test.

*Example*

*Poor:* The (Ordinance) of (1787) provided for the (admission) of (new states).

*Better:* The procedure for admitting new states to the Union was first set forth by the (Ordinance of 1787).

4. *Blanks are Better Put Near the End of a Statement Rather Than at the Beginning.* This permits the problem to be stated before the blank is encountered.

*Example*

*Poor:* A(n) (tariff) is a tax on goods imported into a country.

*Better:* A tax levied on goods imported into a country is called a(n) (tariff).

5. *If the Problem Requires a Numerical Answer, Indicate the Units in Which It Is to Be Expressed.* This will simplify the problem of scoring and will remove one possibility of ambiguity in the examinee's response.

## MULTIPLE-CHOICE ITEMS

The multiple-choice item is the most flexible and most effective of the objective item types. It is effective for measuring information,

vocabulary, understandings, application of principles, or ability to interpret data. In fact, it can be used to test practically any educational objective that can be measured by a pencil-and-paper test except the ability to organize and present material. The versatility and effectiveness of the multiple-choice item is limited only by the ingenuity and talent of the item writer.

The multiple-choice item consists of two parts: the stem, which presents the problem, and the list of possible answers or options. The stem may be presented in the form of an incomplete statement or a question.

### Example

*Incomplete statement:* If both the President and Vice-President died in office, the person who would act as President would be the

- A. Majority Leader of the Senate.
- B. President of the Senate.
- C. Speaker of the House of Representatives.
- D. Secretary of State.

*Question:* Who would act as President if both the President and Vice-President died in office?

- A. The Majority Leader of the Senate.
- B. The President of the Senate.
- C. The Speaker of the House of Representatives.
- D. The Secretary of State.

Inexperienced item writers usually find it easier to use the question form of stem than the incomplete sentence form. The use of the question forces the item writer to state the problem explicitly. It rules out certain types of faults that may creep into the incomplete statement, which we will consider presently. However, the incomplete statement is often more concise and pointed than the question, if it is skillfully used.

The number of options used in the multiple-choice question differs in different tests, and there is no real reason why it cannot vary for items in the same test. However, to reduce the guessing factor, it is preferable to have four or five options for each item. On the other hand, it seems more sensible to have only three good options for an item than to have five, two of which are so obviously wrong that no one ever chooses them.

The difficulty of a multiple-choice item will depend upon both the "closeness" of the options and the process called for in the item.

Consider the set of three items shown below, all relating to the First Amendment to the Constitution. We can predict with some confidence that version I will be passed by more pupils than will II, and II by more than III. The difference between I and II is in the closeness of the options—in I, the wrong choices fall completely outside the Bill of Rights, i.e., the first ten Amendments to the Constitution, while in II, each option refers to some one of these Amendments. The difference between II and III is primarily a matter of the intellectual process involved—II requires little more than remembering and recognizing the key concept involved in the different amendments, while III requires that the student identify that concept when it is embedded in a specific concrete situation.

#### *Version I*

The First Amendment to the Constitution is concerned with

- A. powers of Congress.
- B. the abolition of slavery.
- C. freedom of speech, press, and religion.
- D. the term of office of the President.

#### *Version II*

According to the First Amendment to the Constitution, the government is not permitted to

- A. search a person's house without a warrant.
- B. hold a person in jail for a long time without a trial.
- C. make laws that interfere with freedom of speech or religion.
- D. force a person to give evidence against himself.

#### *Version III*

Which of the following actions would violate the rights guaranteed to a person by the First Amendment to the Constitution?

- A. An F.B.I. agent gets a tip that counterfeiters are operating in Mr. Jones' basement, so he breaks in the door to search the basement.
- B. Mr. Smith is arrested and held in jail for three weeks but is not informed of the charges against him and is not allowed to see a lawyer.
- C. Mr. Simpson is arrested for writing articles criticizing the government's defense policies.
- D. Mr. Hoffman, who is on trial for conspiracy, is forced to take the witness stand and give evidence.



# MAXIMS FOR MULTIPLE-CHOICE ITEMS

1. *The Stem of a Multiple-Choice Item Should Clearly Formulate a Problem.* All the options should be possible answers to a single problem that is raised by the stem. When the stem is phrased as a question, it is clear that a single problem has been raised, but this should be equally the case when the stem is in the form of an incomplete statement. Avoid items that are really a series of unrelated true-false items dealing with the same general topic.

## Example

*Poor:* At the Constitutional Convention, the "great" compromise

- A. gave small and large states equal representation in the Senate.
- B. made slave holding legal.
- C. was opposed by Washington.
- D. gave the western lands claimed by the states to the federal government.

*Better:* At the Constitutional Convention, the "great" compromise between the large and small states was concerned with

- A. representation in Congress.
- B. importation of slaves.
- C. the power to levy taxes.
- D. commerce between states.

2. *Include as Much of the Item as Possible in the Stem.* In the interests of economy of space, economy of reading time, and clear statement of the problem, it is usually desirable to try to word and arrange the item so that the stem is relatively long and the several options relatively short. This cannot always be achieved but is an objective to be worked toward. This principle ties in with the one previously stated of formulating the problem fully in the stem.

## Example

*Poor:* According to the Constitution, neither Congress nor the states can pass a law

- A. that would require a citizen to be able to read and write before he could vote.
- B. that would prevent a citizen from voting because he did not own property.
- C. that would make it impossible for a citizen to vote because he had committed a crime.
- D. that would deprive a citizen of the right to vote because he was of Chinese descent.

*Better:* According to the Constitution, neither Congress nor the states can pass a law that would deprive a citizen of his right to vote because he

- A. could not read or write.
- B. did not own property.
- C. had committed a crime.
- D. was of Chinese descent.

3. *Don't Load the Stem Down with Irrelevant Material.* In certain special cases, the purpose of an item may be to test the examinee's ability to identify and pick out the essential facts. In this case, it is appropriate to hide the crucial aspect of the problem in a set of details that are of no importance. Except for this case, however, the item should be written so as to make the nature of the problem posed as clear as possible. The less irrelevant reading the examinee has to do, the better.

#### Example

*Poor:* The framers of the Constitution faced many problems. The delegates to the Constitutional Convention represented states with different interests, and the delegates from the individual states wanted to see that their states' interests were protected. However, the delegates agreed that the Articles of Confederation needed to be changed in order to provide for

- A. a President whom everyone could respect.
- B. a stronger central government.
- C. a better understanding between states.
- D. a government for and by the people.

*Better:* The delegates to the Constitutional Convention agreed that the Articles of Confederation needed to be changed in order to provide for

- A. a President whom everyone could respect.
- B. a stronger central government.
- C. a better understanding between states.
- D. a government for and by the people.

4. *Be Sure that There Is One and Only One Correct or Clearly Best Answer.* It hardly seems necessary to specify that a multiple-choice item must have one and only one right answer, but in practice this is one of the most pervasive and insidious faults in item writing. Thus, in the following example, though choice A was probably designed to be the correct answer, there is a large element of correctness also in choices B and D. The item could be improved as shown in the revised form.

*Example*

*Poor:* The adoption of the Constitution was generally opposed by people who

- A. owed money.
- B. thought that most of the people were unfit to govern themselves.
- C. owned businesses.
- D. thought the states would be destroyed.

*Better:* The adoption of the Constitution was generally opposed by people who

- A. owed money.
- B. were engaged in commerce.
- C. owned western land.
- D. were engaged in manufacturing.

5. *Items Designed to Measure Understandings, Insights, or Ability to Apply Principles Should Be Presented in Novel Terms.* If the situations used to measure understandings follow very closely the examples used in text or class, the possibility of a correct answer being based on rote memory of what was read or heard is very real. The second and third variations of the example on p. 72 illustrate an attempt to move away from the form in which the concept was originally stated.

6. *Beware of Clang Associations.* If the stem and the keyed answer "sound alike," the examinee may get the question right just by using this superficial cue. However, superficial associations in the wrong answers represent one of the effective devices for attracting those who do not really know the fact or concept being tested. This last practice must be used with discretion, or one may prepare trick questions.

*Example*

*Poor:* A system of checks and balances was established by the Constitution in order to

- A. balance majority power and minority rights.
- B. appease the small states.
- C. distribute powers between the central government and the state governments.
- D. provide for flexibility in the central government.

*Better:* A system of checks and balances was established by the Constitution in order to

- A. prevent one group or one person from seizing the power of government.
- B. balance the powers of the small and large states.
- C. distribute powers equally between the central government and the state governments.
- D. provide for flexibility in the Constitution.

7. *Beware of Irrelevant Grammatical Cues.* Be sure that each option is a grammatically correct completion of the stem. Cues from the use of the indefinite article ("a" versus "an") in the stem, the number or tense of a verb, the use of the plural form of a noun or pronoun, etc., must be excluded.

#### *Example*

*Poor:* A power of the federal government that is suggested by the Constitution but is not directly stated in the Constitution is called an

- A. concurrent power.
- B. residual power.
- C. implied power.
- D. delegated power.

*Better:* A power of the federal government that is suggested by the Constitution but is not directly stated in the Constitution is called

- A. an executive power.
- B. a concurrent power.
- C. an implied power.
- D. a residual power.

(Note that one option was changed to provide for two options that used "an" since test-wise examinees sometimes use the one article that is different as a cue to the correct answer.)

8. *Beware of the Use of One Pair of Opposites as Options If One of the Pair is the Correct or Best Answer.* The directions for a multiple-choice test usually instruct the examinee to choose the one correct or best answer. If only one pair of opposites is used as options and one of the pair is the correct answer, the examinee is likely to limit his choice of answers to these two options because he thinks that both of them cannot be wrong. When this happens, the item is likely to operate as a two-choice item rather than as a four- or five-choice item, and the probability of guessing the correct answer is increased. It is better, if possible, to use two pairs of opposites or to eliminate the use of opposites.

### Example

*Poor:* The chief objective of Daniel Shay's Rebellion was to force the state of Massachusetts to

- A. grant ex-soldiers the right to vote.
- B. issue paper currency.
- C. withdraw paper currency.
- D. stop slave trading.

*Better:* The chief objective of Daniel Shay's Rebellion was to force the state of Massachusetts to

- A. grant ex-soldiers the right to vote.
- B. issue paper currency.
- C. pay ex-soldiers for their services in the Revolutionary War.
- D. stop slave trading.

9. Beware of the Use of "None of These," "None of the Above," "All of These," and "All of the Above" as Options. Except for items requiring numerical computation the option "None of these" or "None of the above" usually fails to make any sense since it contradicts the stem or does not complete the stem grammatically. As a rule both options tend to be used as fillers, i.e., when the item writer cannot think of a fourth or fifth answer choice, he sticks in "None of these" or "All of these" usually as an incorrect answer.

The use of the option "All of these" as a correct answer in a four- or five-choice item generally makes an item less discriminating because if the examinee knows that at least two of the answer choices are correct he automatically gets the correct answer whether he knows anything about the other options or not.

If "None of these" or "All of these" is used as an answer choice, it should be used as frequently for the correct choice as are any of the other options.

### Examples

*Poor:* The Federalist Papers were written by

- A. Hamilton.
- B. Jay.
- C. Madison.
- D. All of the above.

*Better:* The Federalist Papers were written by

- A. Hamilton, Jay, and Madison.
- B. Hamilton, Jefferson, and Madison.
- C. Jefferson, Washington, and Franklin.
- D. Washington, Franklin, and Jay.

*Poor:* Under the Articles of Confederation the national government obtained money to run the government by

- A. putting a tax on imports.
- B. printing additional paper currency.
- C. borrowing money from foreign governments.
- D. none of the above.

*Better:* Under the Articles of Confederation the national government obtained money to run the government by

- A. putting a tax on imports.
- B. printing additional paper currency.
- C. borrowing money from foreign governments.
- D. taxing property.

10. *Use the Negative Only Sparingly in the Stem of an Item.* It is usually desirable to emphasize the positive aspects of knowledge rather than the negative aspects of knowledge. However, there are times when it is important for the student to know the exception or to be able to detect errors. For these purposes a few items with the words "not" or "except" in the stem may be justified, particularly when over-inclusion is a common error for students. When a negative word is used in the stem of an item, it should be underlined and/or capitalized to call the student's attention to it.

### *Example*

*Poor:* Which one of the following leaders of the Revolutionary War did not want the Articles of Confederation changed?

- A. Benjamin Franklin
- B. George Washington
- C. Alexander Hamilton
- D. Patrick Henry

(Note this is a poor use of the negative stem because it could be stated more effectively in positive form, "Which one of the following leaders of the Revolutionary War favored keeping the Articles of Confederation?")

*Better:* According to the Constitution, the President does NOT have the power to

- A. declare war.
- B. pardon a person convicted by a federal court.
- C. call a special session of Congress.
- D. nominate judges for the Supreme Court.

(Note this is a better use of the negative stem because (1) it requires the student to detect a common error made about the powers of the President; and (2) it would be difficult to get three good misleads if the item were stated in positive form. The stem of the item could not be stated "The Constitution forbids the President to \_\_\_\_\_" because the Constitution does not specifically forbid the President to declare war.)

# THE MATCHING ITEM

The matching item is actually a special form of the multiple-choice item. The characteristic that distinguishes it from the ordinary multiple-choice item is that instead of a single problem or stem with a group of suggested answers, there are several problems whose answers must be drawn from a single list of possible answers.

The matching item has most frequently been used to measure factual information such as the meaning of words, dates of events, association of authors with titles of books or titles with plot or characters, names associated with particular events, or association of chemical symbols with names of chemicals. The matching item is a compact and efficient way of measuring this type of achievement.

Effective matching items may often be built by basing the set of items upon a graph, chart, map, diagram, or picture of equipment. Features of the figure may be labeled, and the examinee may be asked to match names, functions, etc., with the labels on the figure. This type of item is particularly useful in tests dealing with science or technology, e.g., identification of organs in an anatomy test.

However, there are many topics to which the matching item is not very well adapted. The items making up a set should bear some relationship to each other; that is, they should be homogeneous. In the case of many of the outcomes one would like to test, it is difficult to get enough homogeneous items to make up a set for a matching item.

Consider the example that appears below.

*Instructions:* Match the statements in Column I with those in Column II.

Column I	Column II
1. First Ten Amendments	A. 1215
2. We owe much of our democratic heritage to this country.	B. George Washington
3. Date of the Magna Charta	C. Authors of the <i>Federalist Papers</i>
4. Jay, Hamilton, and Madison	D. England
5. Chairman of the Constitutional Convention	E. Bill of Rights

This example illustrates most of the common mistakes made in preparing matching items. First, the directions are vague because they do not specify either the basis for matching or how the examinee is to record his answers. Second, the statements in Column I have nothing in common except that all of them refer to materials usually included in an eighth-grade unit on the Constitution. Look at statement 3 in Column I which asks for the date of the Magna Charta. Column II includes only one date. Successful matching here requires no knowledge on the part of the student. Each item in the set can

be matched in the same way, using only the most superficial cues. Third, note the number of answer choices provided in Column II to match with the five statements in Column I. If the instructions indicate that each answer is to be used only once, then the person who knows four of the answers automatically gets the fifth by elimination, and the person who knows three has a fifty-fifty chance on the last two.

#### MAXIMS ON MATCHING ITEMS

1. *When Writing Matching Items, the Items in a Set Should Be Homogeneous.* For example, they should all be names of persons, or all dates of events, or all provisions of different parts of the Constitution.

2. *The Number of Answer Choices Should Be Greater Than the Number of Problems Presented.* This holds except when each answer choice may be used more than once, as in variations that we shall consider presently.

3. *The Set of Items Should Be Relatively Short.* It is better to make several relatively short matching sets than one long one because (1) it is easier to keep the items in the set homogeneous and (2) it is easier for the student to find and record the answer.

4. *Response Options Should Be Arranged in a Logical Order, if One Exists.* Arranging names in alphabetical order or dates in chronological order reduces the clerical task for the examinee.

5. *The Directions Should Specify the Basis for Matching and Should Indicate Whether an Answer Choice May Be Used More Than Once.* These precautions will guarantee a more uniform task for all examinees.

A variation on the matching type of item which is sometimes effective is the classification type or master list. This pattern, illustrated on p. 81, presents an efficient means of exploring range of mastery of a concept or related set of concepts.

Another setting in which the master list variation of the classification type of item can often be used to advantage is that of testing knowledge of the general chronology or sequence of events. See Example II on p. 81.

There are a number of other varieties of objective test items that have been developed and used to some extent. The reader who is interested in a survey of these, together with a more extended discussion of teacher-made tests in general, is referred to the suggested additional readings at the end of the chapter.



*Example I*

*Instructions:* Below are given some happenings that could take place in a session of Congress. For each of these, you are to mark

- A. if it is specifically permitted or required by the Constitution.  
 B. if it is specifically forbidden by the Constitution.  
 C. if it is implied by the Constitution but nothing is specifically stated about it.  
 D. if it has developed through custom and usage.
- (A) 1. At the opening session of Congress, the President delivers a "State of the Union Address."  
(C) 2. Congress passes a law raising minimum wages from \$1.00 to \$1.25.  
(B) 3. Congress passes a law requiring all children to attend school until they reach the age of 16.  
(D) 4. The President requests a senator from State X to suggest names of persons who would be satisfactory as collectors of customs.  
 (and possibly others).

*Example II*

For each event on the left, pick the choice on the right that tells when the event took place.

<i>Events</i>	<i>Time Line</i>
<u>(E)</u> 1. Women were granted suffrage.	A Declaration of Independence.
<u>(D)</u> 2. All persons born or naturalized in the U. S. were declared to be citizens.	B Adoption of the Constitution.
<u>(A)</u> 3. Zenger trial was held.	C Civil War.
<u>(B)</u> 4. Plan for admitting new states was adopted.	D World War I.
<u>(C)</u> 5. Freedom of religion, speech, and press were guaranteed (and possibly others).	E

## TESTING FOR UNDERSTANDING

Since it is easier to construct questions testing factual knowledge than those that measure understanding, application of principles, and other meaningful outcomes of instruction, teacher-made tests, espe-

cially of the objective type, tend to emphasize facts. Teachers tend to assume that if a student knows the factual material, then he also understands that material. Although there is a positive relationship between factual knowledge and understanding, the relationship is not perfect. It is true that in order for the student to understand a principle, he must have the relevant facts and basic skills. But there is no assurance that mere possession of the facts means that the student really understands the material.

If students are to develop understandings, understandings must be taught and they must be evaluated. In the measurement of understanding, the situations or applications used in evaluation should be similar to, but not identical with, the examples used in class. If the same situations are used, the student may get the correct answer because he has memorized the example given in class, not because he understands the principle.

Objective test items do not divide up into two clearly distinct groups, those that measure factual knowledge and those that measure understanding, application, or interpretation. Many items involve understanding and application at various levels as well as the underlying factual knowledge. Thus, illustration III on p. 72 and the matching item on p. 81 both call for applications of knowledge to new situations. Multiple-choice items in particular readily lend themselves to testing the understanding and application of principles with novel material or in novel settings.

Another type of item is the interpretive type item. This type of item consists of an introductory selection of material, giving the necessary background and setting the problem, followed by a series of questions asking for interpretations of the material. The introductory material can be text, graphs, tables, maps, charts, or any similar material. It can be complete in itself, providing all the necessary information basic to the understanding, or it can be incomplete so that the student must know certain things in addition to those given.

The eighth-grade unit on the Constitution that we have used so far does not provide for good examples of the interpretative type of exercise. However, two examples of the interpretive test exercise that were constructed for a twelfth-grade unit on labor unions are given. The first is based on a graph showing certain data about union membership, strikes, and important social and economic events. In this item, the accuracy of the student's answer depends only upon his ability to understand the material as it is presented to him in graphic form.

The second example is based on a newspaper item, and the student

is not given all the essential information but must know certain facts about the Taft-Hartley Act in order to answer the question.

### Example 1

The following statements refer to Fig. 4.1. Read each statement carefully. In front of each statement mark

- A if the statement is supported by the evidence in Fig. 4.1.  
 B if the statement is contradicted by the evidence in Fig. 4.1.  
 C if the statement is neither supported nor contradicted by the evidence in Fig. 4.1.

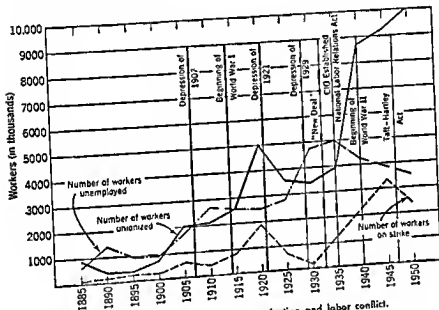


Fig. 4.1. Factors relating to labor organization and labor conflict.

- (B) 1. Bad economic conditions tend to produce large numbers of strikes.  
 (C) 2. The "New Deal" encouraged workers to join unions.  
 (A) 3. The number of workers out on strike increases after a war.  
 (B) 4. The passage of the Taft-Hartley Act caused a drop in union membership.  
 (C) 5. By 1950, the majority of skilled and semiskilled workers in industry belonged to unions.  
 (B) 6. As the number of unemployed workers increases, membership in unions increases.  
 (C) 7. The large number of men on strike in 1945 caused the passage of the Taft-Hartley Act in 1947.

- (A) 8. The period between 1920 and 1930 was marked by a steady decrease in union membership.
- (A) 9. The establishment of the CIO was followed by an increase in union membership.
- (A) 10. The pattern of number of workers out on strike is similar to that of the number of workers belonging to unions.

### Example II

Radio station WKRX uses only recorded music on its programs. A contract between the radio station and the musicians' union had required the station to hire a certain number of musicians, even though the musicians never played on any programs. When the contract ended, the radio station refused to renew it. Members of the musicians' union started to picket the radio station headquarters to force it to renew the contract. When the baseball season started, members of the union began to picket the local baseball park because the ball games of the local team were broadcast over station WKRX. The owners of the ball team and of the radio station took the case to court and asked the court to rule whether the picketing was legal.

What was the most probable ruling by the court?

- A. Only the picket line at the baseball park was legal.
- B. Only the picket line at the radio station was legal.
- C. Both picket lines were legal.
- D. Neither of the picket lines was legal.

From the statements below check all that support your answer.

- \_\_\_\_\_ 1. Workers cannot be prevented by management from using any peaceful method of protecting their jobs.
- \_\_\_\_\_ 2. The Taft-Hartley Act permits strikes when other means of settling disputes fail.
- X 3. Secondary boycotts are forbidden by the Taft-Hartley Act.
- X 4. "Featherbedding" practices by unions are forbidden under the Taft-Hartley Act.
- \_\_\_\_\_ 5. Since the picketing of the baseball park was against the radio station and not against the baseball team, the owners of the baseball team had no grounds for court action.
- \_\_\_\_\_ 6. Since baseball is a sport, not a business, a baseball park cannot be used to force the settlement of a dispute between labor and management.
- \_\_\_\_\_ 7. Strikes cannot be called against an employer who does not have a contract with a union.

The interpretive type of item provides an opportunity to ask meaningful questions about complex data in order to evaluate the student's ability to understand and interpret such materials.

However, this item type presents special problems. The introductory material must be carefully chosen to elicit the type of understanding that the teacher desires. Although a number of sources such as

newspapers, magazines, or books can be used to furnish the introductory material, it usually has to be rewritten and adapted by the teacher to keep it at an appropriate reading level and to eliminate unnecessary parts. The success of this type of item is dependent to a large extent upon the adequacy of the introductory material.

Another disadvantage of the interpretive type of item is the reading load. Most of these items tend to be long, so that the evaluation of understanding will be contaminated by the reading level of the student.

A third disadvantage is the amount of space required to present the item and the amount of time required to answer it. With this type of item it is not possible to get as many different units of coverage as with the usual type of multiple-choice item.

For a more detailed discussion and for more examples of methods of measuring understanding in the different subject-matter fields, the reader is referred to the *Forty-Fifth Yearbook of the National Society for the Study of Education*, listed in the supplementary readings at the end of the chapter.

## GETTING THE OBJECTIVE TEST READY FOR USE

So far we have considered the problems involved in improving the quality of the individual objective test items. Now we must give some thought to putting the items together into a test that is an effective whole. The quality of the total test will have been determined in large measure by the quality of our initial planning and by the skill with which we have written the separate test items. However, some further suggestions may help in achieving a sound and workmanlike product.

### EXTRA ITEMS

When the items are originally written it will usually pay to write a surplus over the number that will finally be used. Items that seem masterworks in the first pride of authorship may show unsuspected flaws when coldly re-examined at a later date. Furthermore, some freedom for fitting the final test to the specifications of the blueprint is often helpful. A surplus of 20 or 30 per cent is none too much.

### REVIEW AND EDITING

It is always sound policy, if time permits, to write the items early and put them aside for a while. When reread later, ambiguities will appear that were not seen at all when the item was first written. Even more helpful, if it is feasible, is to get another person who knows the subject matter to go over the items, keying them and criticizing them. This type of review will usually bring out a rather startling number

of points of ambiguity or disagreement. Revision of the items in the light of such a critique or elimination of items that seem not to be salvageable will do much to avoid those debates with students and those ill-feelings that are an occasional feature of objective examinations.

#### FORM OF REPRODUCTION

Though it is possible to give objective examinations orally, it is far from satisfactory to do so. Oral administration is demanding upon students' concentration and introduces an element of speed pressure that is quite disturbing to some. One generally assumes that an objective test will be reproduced and that each pupil will have a copy. Gelatin duplicating processes are adequate for groups of moderate size, but most test makers will prefer to mimeograph the test if facilities for mimeographing are available. More important than the process is the quality of the work, both in organizing the layout of the test and in typing up the master copy.

#### ORDER AND GROUPING OF TEST ITEMS

After the items have been edited and those to be included in the test have finally been selected, they must be arranged in the order in which they are to appear in the test. There are three aspects that should be considered and reconciled as far as possible in deciding upon the arrangement and grouping of items.

1. Items in the same format (true-false, multiple-choice, etc.) should be grouped together, so that instructions for answering will carry throughout the set.

2. In general, an attempt should be made to progress from easy to more difficult items. This is especially important with younger children, who may become discouraged and quit if the early items are too difficult. It is also important if time is likely to be limited, so that some items will not be reached. These not-attempted items should be the more difficult ones that the examinee would not have been likely to answer correctly even if he had reached them.

3. Items dealing with similar content can well be grouped together. If this is done, it will help to reduce the feeling that the test is made up of unrelated bits and pieces. It will encourage a more integrated attack by the examinee.

#### DIRECTIONS

Clear instructions to the examinees are an important element in a well-constructed test. Examinees will usually know the purpose of a test, but if it is possible that they may not the purpose should be

stated. Complete instructions should be given as to how the pupil is to record his answers. This is particularly important for novel or unusual item patterns. The examinee should be given explicit information as to the scoring procedure that will be used, including the credit for each item or part and whether or not a correction will be made for guessing. (See *Scoring*, p. 88.)

Sample sets of directions for matching items have been given on p. 81.

For a test made up of multiple-choice items that will not be corrected for guessing and for which separate answer sheets are used, one might use the following set of directions.

*Directions:*

Read each item and decide which choice *best* completes the statement or answers the question.

Mark your answers on the separate answer sheet. Do not mark them on the test booklet. Indicate your answer by blacking out on the answer sheet the letter corresponding to your choice. That is, if you think that choice B is the best answer to item 1, black out the B in the row after No. 1 on your answer sheet.

Your score will be the number of right answers, so it will be to your advantage to answer every question, even if you are not sure of the right answer.

*Be sure your name is on your answer sheet.*

For a test made up of true-false questions in which answers are to be recorded on the test paper and the total score will be corrected for guessing, the following set of directions could be used.

*Directions:*

Read each of the following statements carefully.

If all or any part of the statement is false, circle the F in front of the statement.

If the statement is completely true, circle the T in front of the statement.

Your score will be the number of right answers minus the number of wrong answers, so do not guess blindly. If you are not reasonably sure of an answer, omit the question.

*Be sure your name is on your test.*

#### LAYOUT OF ITEMS

The two points important to bear in mind when planning how the items and answers will be placed on the sheet are (1) clarity and convenience for the examinee and (2) convenience for the scorer. In the interest of the person taking the test, items should not be crowded together too closely. Multiple-choice items are easier to read if each response option is on a separate line. Having part of an item on one

Course _____	Name _____
Exam _____	Date _____

Instructions: Read the directions on the test sheet carefully, and follow them exactly. For each test item, mark your choice for the correct answer by blocking out the letter which corresponds to the best answer for the test item.

Item	Answer	Item	Answer	Item	Answer
1	A B C D E	26	A B C D E	51	A B C D E
2	A B C D E	27	A B C D E	52	A B C D E
3	A B C D E	28	A B C D E	53	A B C D E
4	A B C D E	29	A B C D E	54	A B C D E

Fig. 4.2. Part of a home-made answer sheet.

page and part on the next should be avoided if possible. If several items all refer to a single diagram or chart, it is desirable that all of them appear on the same page as the diagram or chart.

The arrangement of answers should be such as to facilitate scoring. Even in the upper elementary school it is practical to put spaces for all the answers in a column on one side of the page. A scoring key can then be laid beside the answer column to speed up scoring. In the junior high school and above, a simple separate answer sheet may be used. Part of a home-made answer sheet which is adaptable for both true-false and multiple-choice items is shown in Fig. 4.2.

In school-wide or city-wide testing projects, machine-scored answer sheets of the type developed for standardized tests may be used, if facilities for machine scoring are available.

### SCORING

Layout of answers to facilitate scoring has been discussed in the previous paragraphs. A scoring stencil that can be placed alongside the columns of answers or placed directly over a separate answer sheet will make scoring go very quickly.

The test maker must decide how he is going to treat guessing in his scoring procedure. As we have indicated, his decision should be made known to the examinees. If time permits every student to at-



tempt every item, a score that is simply the number of right answers is quite satisfactory. In this case, examinees should be firmly instructed to guess, even if they have no idea of the answer. This procedure has sometimes been criticized as poor pedagogy, since it involves practice in errors. However, the student will think about each item anyhow. It seems doubtful that the final step of marking an answer, when one knows in one's own mind that one is just guessing, will have any very lasting impact on the impression one carries away from the test.

If the test is speeded, so that pupils will attempt different numbers of items, or if the test user wishes to discourage guessing on the part of examinees, a penalty should be applied for wrong answers. The usual correction formula, based on the assumption that the person who does not know the answer will make a *random guess*, is

$$\text{Score} = R - \frac{W}{n - 1}$$

where  $R$  is the number of questions answered correctly;  
 $W$  is the number of questions answered incorrectly;  
 $n$  is the number of answer choices for an item.

For example, in a true-false test where there are only 2 possible answers,  $n - 1$  becomes  $2 - 1$ , or 1, and the correction for guessing is the number of right answers minus the number of wrong answers. Thus, if there were 75 true-false items on a test and a student got 48 right, got 20 wrong, and did not answer 7 of them, his score would be  $48 - 20$  or 28. Note that omissions do not count in this formula for guessing.

For a second example, suppose a student took a 60-item multiple-choice test in which each item had 5 possible answers. If he got 52 questions right and 8 wrong, his corrected score would be

$$52 - \frac{8}{5 - 1} \quad \text{or} \quad 52 - \frac{8}{4} = 50$$

## ANALYZING AND USING THE RESULTS OF OBJECTIVE TESTS

Giving the test, scoring it, and recording a score for each pupil frequently ends the matter as far as the teacher is concerned. However, if the teacher drops the test at this point, he loses much of its value. An analysis of the responses the pupils made to the items can serve two important purposes. In the first place, the test results pro-

vide a diagnostic technique for studying the learnings of the class and the failures to learn and for guiding further teaching and study. In the second place, the responses of pupils to the separate items and a review of the items in the light of these responses provide a basis for preparing better tests another year.

The basic analysis that is needed is a tabulation of the responses that have been made to each item on the test. We need to know how many pupils got each item right, how many chose each of the possible wrong answers, and how many omitted the item. It helps our understanding of the item if we have this information for the upper and lower fractions of the group, and perhaps also for those in the middle. From this type of tabulation, we can answer such questions as the following for each item:

1. How hard is the item?
2. Does it distinguish between the better and poorer students?
3. Do all the options attract responses, or are there some that are so unattractive that they might as well not be included?

A simple form can be prepared for recording the responses to each item, like that shown in Fig. 4.3. This can be put on a separate card for each item, and then the information can be accumulated in a permanent item file. This form is planned for a multiple-choice item with as many as five choices but can be used for true-false items by using only the A and B columns.

*Item:* Which one of the following states was formed from the Northwest Territory?

- A. Indiana  
B. Iowa  
C. Montana  
D. Oregon

	Option					
	A	B	C	D	E	Omit
Upper 25%	10					
Middle 50%	17		1	2		
Lower 25%	5	1	1	3		

Fig. 4.3. Form for recording item-analysis data.

To illustrate the type of information that is provided by an item analysis, we present below certain items from a social studies test, together with the analysis of responses for each item. This test was given in 1960 to 100 high-school seniors who had had a course in current American problems. There were 95 items on the test. The highest score on the test was 85 and the lowest score was 14. The test papers were arranged in order of total score starting with the score of 85 and ending with the score of 14. The top 25 papers were selected to represent the upper group (score range 59 to 85) and the last 25 papers were selected to represent the lower group (score range 14 to 34). The count of responses is based on the 25 cases from the top and the 25 cases from the bottom of the group. The responses made to each item by each individual in the upper and lower groups were tallied to give the frequency of choosing each option. These frequencies are shown on the right. The correct option is underlined. Each item is followed by a brief discussion of the item data.

#### Item I

"Everyone's switching to Breath of Spring Cigarettes!" is an example of the propaganda technique called

	Upper	Lower
A. glittering generality.	0	2
<u>B. bandwagon.</u>	25	20
C. testimonial.	0	2
D. plain folk.	0	1
(Omit)	0	0

This is an easy item, since all 25 in the upper group and 20 in the lower group get it right. However, it does differentiate in the desired direction, since what errors there are fall in the lower group. The item is also good in that all of the wrong answer choices are functioning; i.e., each wrong answer has been chosen by one or more persons in the lower group. Two or three easy items like this would be good "ice-breakers" with which to start a test.

#### Item II

There were no federal income taxes before 1913 because prior to 1913

	Upper	Lower
A. the federal budget was balanced.	3	5
B. regular property taxes provided enough revenue to run the government.	9	15

C. a tax on income was unconstitutional.	13	0
D. the income of the average worker in the U. S. was too low to be taxed.	0	5
(Omit)	0	0

This was a difficult item but a very effective one. That it was difficult is shown by the fact that only 13 out of 50 got it right. That it was effective is shown by the fact that all 13 getting the item right were in the upper group. All of the wrong options attracted some choices in the lower group and all of the wrong options attracted more of the lower group than the higher group. Incidentally, an item such as this shows how faulty the idea of "blind guessing" often is when an item is effectively written. In this item, the majority of the lower group concentrated upon one particular wrong option that was particularly plausible and appealing.

### Item III

Under the "corrupt practices act" the national committee of a political party would be permitted to accept a contribution of

	Upper	Lower
A. \$10,000 from Mr. Jones.	15	4
B. \$1,000 from the ABC Hat Corporation.	4	6
C. \$5,000 from the National Association of Manufacturers.	2	8
D. \$500 from union funds of a local labor union.	4	7
(Omit)	0	0

This item turned out poorly. Only 10 out of 50 got it right, and right answers were more frequent in the lower than in the upper group. As far as the test is concerned, it appears that this item would have to be either discarded or radically revised. If the group was supposed to have learned about the provisions of the "corrupt practices act," this shows clearly that the learning did not take place. In order to arrive at the correct answer to the item the student would have to know (1) the limit placed on contributions to the national committee of a political party, (2) who is forbidden to make contributions, and (3) what kind of organization the National Association of Manufacturers is. The teacher would have to discuss the item with the class to determine where the difficulty lies but one might guess that it is points 1 and 3 that are causing difficulty in the upper group.

*Item IV*

The term "easy money" as used in economics means

	<i>Upper</i>	<i>Lower</i>
<u>A.</u> the ability to borrow money at low interest rates	21	17
B. dividends that are paid on common stocks.	0	0
C. money that is won in contests.	0	0
D. money paid for unemployment compensation.	4	8
(Omit)	0	0

This item shows some discrimination in the desired direction (21 versus 17), but the differentiation is not very sharp. The response pattern is one that is quite common. Only two of the four choices are functioning at all. Nobody selects either the B or C choices. If we wished to use this item again, we might try substituting "wages paid for easy work" for option B and "Money given to people on welfare" for option C. The repeat of the word "easy" in option B and the idea of getting money for not working in option C might make the item more difficult and more discriminating.

Item statistics such as these can be used not only for evaluating the items but to guide review and restudy of the material with a class. The items that prove difficult for the class as a whole provide leads for further exploration. Discussion of these items with the class should throw light on the nature of the misunderstanding. The misunderstanding may in some cases be cleared up by brief further discussion, although in some cases a fuller review of the topic may be indicated. It is desirable, if local policies permit, to let pupils have their answer sheets and a copy of the test and to make the answer key available to them, so that they can themselves use the test as a guide to review and clarification of the points they missed. An examination should teach as well as test.

## SUMMARY STATEMENT

The deficiencies of essay examinations have led to the preparation of tests made up of objective short-answer questions. These questions may be prepared in true-false, completion, multiple-choice, matching, and many other forms. Experience of item writers has led to the formulation of a number of "do's" and "don't's" to guide the preparation of test items. These are considered in detail in this chapter.

Though there is an unfortunate tendency for writers of objective items to concentrate on factual information, ability to understand, interpret, and apply can be tested by items that follow this format. For the measurement of understanding it is often desirable to describe a fairly complex problem situation or to present a fairly full set of data and to organize a set of related questions about the problem or data. Illustrations are provided.

It helps, in producing a good test, to prepare extra items and to have the items edited and screened before using. Items should be grouped so as to emphasize relationships and to provide a general progression from easy to more difficult. Answer sheets and scoring stencils facilitate scoring. The issue of correction for guessing should be resolved in advance, and examinees should be told what procedure will apply.

Test results can be analyzed with profit to guide (1) further teaching and review and (2) the construction of additional tests in later years.

### SUGGESTED ADDITIONAL READING

- Dressel, Paul L., and Lewis B. Mayhew, *Science reasoning and understanding*, Dubuque, Iowa, William C. Brown, 1954.
- Ebel, Robert L., Writing the test item, Chapter 7 in E. F. Lindquist, Editor, *Educational measurement*, Washington, D. C., American Council on Education, 1951.
- Gerberich, J. Raymond, *Specimen objective test items*, New York, Longmans, Green, 1956.
- Micheels, William J., and M. Ray Karnes, *Measuring educational achievement*, New York, McGraw-Hill, 1950.
- National Society for the Study of Education, *The measurement of understanding*, The Forty-Fifth Yearbook, Part I, Chicago, Illinois, University of Chicago Press, 1946.
- Odell, C. W., *How to improve classroom testing*, rev. ed., Dubuque, Iowa, William C. Brown, 1958, chapters VII-XIII.
- Traxler, Arthur E., Administering and scoring the objective test, Chapter 10 in E. F. Lindquist, Editor, *Educational measurement*, Washington, D. C., American Council on Education, 1951.
- Wood, Dorothy Adkins, *Test construction, development and interpretation of achievement tests*, Columbus, Ohio, Charles E. Merrill, 1960.

### QUESTIONS FOR DISCUSSION

1. A high-school principal has a system of using a different type of objective test item each month—one month it is true-false, the next month multiple-choice, the next month completion, and so on. Each teacher is expected to follow this uniform pattern. How would you evaluate this procedure? Why?

2. What steps can a teacher take to avoid ambiguous items on an objective test?

3. Under what conditions would it be important to correct scores on an objective test for guessing?

4. Collect some examples of poor items you have seen on tests. Indicate what is wrong with each item.

5. Construct four multiple-choice items designed to measure understanding or application in some subject area in which you are interested.

6. Prepare a short objective test for a small unit that you are teaching or plan to teach. Indicate the objectives that you are trying to evaluate with each item. (Use the blueprint from Question 3, p. 58 if one is available.)

7. What are the arguments for and against returning major examination papers to students?

8. A fourth-grade teacher has given a test in arithmetic. What analyses of the results could the teacher make that would help guide (a) future work for the class as a whole and (b) special assistance given to individual pupils?

9. A college teacher has given an objective test to a large class, scored the papers, and entered the scores in the class record book. What further steps might the teacher take before returning the papers to the students? Why?

## Chapter 5



# Elementary Statistical Concepts

### INTRODUCTION

In its various forms, measurement results in classification, rankings, or scores. Any attempt to describe, summarize, or compare results for individuals or for groups calls for *numerical* treatment. The branch of arithmetic and mathematics that deals with the analysis of sets of scores for groups of individuals is known as statistics. Every user of tests and measurement devices needs at least a consumer's understanding of the basic objectives and techniques of descriptive statistics. This is a book on measurement, not a statistics textbook. Discussion of statistics as such is limited to this one chapter. It cannot be expected that study of it will make the reader an accomplished statistician. This chapter points out to the novice some basic types of questions that the statistician tries to answer, and introduces him to the simplest tools used to answer them.

Suppose you have prepared tests in reading, arithmetic, and spelling and given them to the pupils in two sixth grades in your school. You have scored the papers and entered the names and scores on a record sheet for the two classes. Table 5.1 shows the way the record

Table 5.1. Record Sheet for Sixth Grades of School X

Name	Test Scores		
	Reading	Arithmetic	Spelling
1. Carol A.	32	3	26
2. Mary B.	27	27	23
3. Ruby C.	31	9	29
4. Alice D.	36	18	27
5. Theresa E.	47	21	35
6. Ida F.	42	24	26
7. Vivian G.	22	4	17
8. Grace H.	50	42	32



Table 5.1. (Continued)

Name	Test Scores		
	Reading	Arithmetic	Spelling
9. Opal I.	20	18	11
10. Ursula J.	37	2	29
11. Beatrice K.	25	10	15
12. Karen L.	37	13	23
13. Susan M.	28	20	25
14. Jane N.	34	15	30
15. Dorothy O.	31	19	22
16. Frances P.	21	2	17
17. Elizabeth Q.	35	48	23
18. Pearl R.	59	41	33
19. Joan S.	44	41	29
20. Nancy T.	32	40	18
21. Judith U.	56	24	39
22. Edith V.	38	24	21
23. Louise W.	38	18	29
24. Helen X.	29	12	27
25. Martha Y.	24	26	22
26. Doris Z.	36	12	30
27. James A.	36	29	25
28. Albert B.	21	16	14
29. Donald C.	27	7	16
30. Peter D.	37	29	21
31. Samuel E.	46	36	32
32. George F.	33	10	27
33. Roger G.	17	14	17
34. Newton H.	35	18	29
35. Karl I.	30	12	19
36. Isidore J.	22	30	12
37. John K.	43	9	33
38. Benjamin L.	31	15	20
39. Theodore M.	50	38	30
40. Michael N.	34	20	20
41. Herman O.	30	15	19
42. Charles P.	52	39	36
43. Patrick Q.	40	33	26
44. William R.	42	6	32
45. Martin S.	17	26	11
46. Frank T.	32	20	18
47. Ralph U.	38	20	22
48. Thomas V.	29	29	24
49. Henry W.	36	25	27
50. Oscar X.	43	19	33
51. Edward Y.	27	19	24
52. Leonard Z.	39	19	25

sheet might look. Now, what sorts of questions might you ask these data? That is, to what questions might you ask the data to provide the answers? Before reading further, suppose you study the set of scores and jot down on a piece of scrap paper the questions that come to *your* mind in connection with these scores. See how many of the question types you can anticipate.

---

A first, rather general question you might ask is: What is the general pattern of the set of scores? How do they "run"? What do they "look like"? How can we picture the set of reading scores, for example, so that we can get an impression of the group as a whole? To answer this question we will need to consider simple ways of tabulating and graphing a set of scores.

A second type of question that will almost certainly arise is: What is this group like, on the average? Have they done as well on the test as other sixth-grade groups? Are they ready for the regular sixth-grade instruction and materials? What is the typical level of performance in the group? All these questions call for some single score to represent the group as a whole, some measure of the middle of the group. To answer this question we shall need to become acquainted with statistics developed to represent the average or typical score.

Third, in order to describe your group you might feel a need to describe the extent to which the scores spread out away from the average value. Are all the children in the group about the same, so that the same materials and procedures would be suitable for all? If not, how widely do they spread out on a given test? How does this group compare with other classes with respect to the *spread* of scores? This calls for a study of measures of variability.

Fourth, you might ask how a particular individual stands on some one test. Thus, you might want to know whether James A. had done well or poorly on the arithmetic test, and if you decided that his score was a good score you might want some way of saying just how good it was. You might ask whether James A. did better in reading or in arithmetic. To answer this question we need a common yardstick in terms of which to express performance in two quite different areas. Our need, then, is for some uniform way of expressing and interpreting the performance of an individual. How does he stand, relative to his group?

A fifth query is of this type: To what extent did those who excelled in reading also excel in arithmetic? To what extent do these two abilities go together in the same individuals? Is the individual who is

superior in one likely also to be superior in the other? To measure this going-togetherness we shall need to become acquainted with measures of *correlation*.

The following sections of this chapter will be devoted to illustrating and discussing the routines that statistics has developed for answering these questions. There are many other questions that may arise with respect to a set of data. The most important ones concern the drawing of general conclusions from data on a limited group. Thus, one sample of fifty boys may have surpassed a sample of fifty girls from the same school on a history test. This is a *descriptive* fact true of these particular groups. We would like to know whether we can safely conclude that the *total population of boys* from which this sample was drawn would surpass the *total population of girls* on this same test. This is a problem of *inference*. Problems of statistical inference make up the bulk of advanced statistical work, but we cannot go into them here.

## WAYS OF TABULATING AND PICTURING A SET OF SCORES

In Table 5.1 we showed a record sheet on which test scores for 52 sixth-grade pupils had been recorded. Let us look at the scores in the column headed Reading and consider how they could be rearranged so as to give us a clearer picture of how the pupils did on the reading test.

The simplest rearrangement would be merely to arrange them in order from highest to lowest. We would then have something that looked like this:

59	43	37	34	30	22
56	42	37	33	29	22
52	42	36	32	29	21
50	40	36	32	28	21
50	39	36	32	27	20
47	38	36	31	27	17
46	38	35	31	27	17
44	38	35	31	25	
43	37	34	30	24	

This arrangement gives a somewhat better picture of the way the scores fall. We can see the highest and lowest scores at a glance, i.e., 59 and 17. It is also easy to see that the middle person in the group falls somewhere in the mid-thirties. We can see by inspection that

roughly half the scores fall between 30 and 40. But this simple rearrangement of scores still has too much detail for us to see the general pattern clearly. It is also not a convenient form to use in computing. We need to condense it into a more compact form.

#### PREPARING A FREQUENCY DISTRIBUTION

A further step in organizing the scores for presentation is to prepare what is termed a *frequency distribution*. This is a table showing how often each score occurred. Each score value is listed, and the number of times it occurred is shown. A portion of the frequency distribution for the reading scores is shown in Table 5.2. However,

Table 5.2. Frequency Distribution of Reading Scores  
(Ungrouped Data)

Test Score	Frequency
59	1
58	0
57	0
56	1
55	0
54	0
53	0
52	1
51	0
50	2
.	.
.	.
.	.
20	1
19	0
18	0
17	2

Table 5.2 is still not a very good form for reporting our facts. The table is too long and spread out. We have shown only part of it. The whole table would take 43 lines. It would have a number of zero entries. There would be *marked ups* and *downs* from one score to the next.

In order to improve the form of presentation further, scores are often *grouped* together into broader categories. In our example, we will group together three adjacent scores, so that each grouping includes three points of score. When we do this, our set of scores is

represented as shown in Table 5.3. This provides a fairly compact table showing how many scores there are in each group or *class interval*. Thus, we have eight scores in the interval 34-36. We do not know how many of them are 34's, how many 35's, and how many 36's. We have lost this information in the grouping. We assume that they are evenly divided. In most cases, there is no reason to anticipate that any one score will occur more often than any other and this assumption is a sound one, so the gains in compactness and convenience of presentation more than make up for any slight inaccuracy introduced by this grouping.\*

Table 5.3. Frequency Distribution of Reading Scores  
(Grouped Data)

Score Interval	Tallies	Frequency
58-60	/	1
55-57	/	1
52-54	/	1
49-51	//	2
46-48	//	2
43-45	///	3
40-42	///	3
37-39	///	7
34-36	///	8
31-33	///	7
28-30	///	5
25-27	///	4
22-24	///	3
19-21	///	3
16-18	//	2

In a practical situation, we always face the problem of deciding how broad the groupings should be, i.e., whether to group by 3's, 5's, 10's, or some other grouping. The decision is a compromise between losing detail from our data, on the one hand, and obtaining a convenient, compact, and smooth representation of our results, on the other. A broader interval loses more detail but condenses the data into a more compact picture. A practical rule-of-thumb is to choose a class interval that will divide the total score range into roughly 15 groups.

\* In some special types of social statistics, such as reports of income, certain values are more likely than others, i.e., \$2000, \$3000, \$5000, etc. Special precautions are necessary in grouping material of this type. In particular, one should strive to get popular values near the *middle* of a class interval.

Thus, in our example the highest score was 59 and the lowest was 17. The range of scores is  $59 - 17 = 42$ . Dividing 42 by 15, we get 2.8. The nearest whole number is 3, and so we group our data by 3's. In addition to the "rule of 15," we also find that intervals of 5, 10, and multiples of 10 make convenient groupings. Since the purpose of grouping scores is to make a convenient representation, factors of convenience enter as a major consideration.

It should be noted that sometimes there is no need to group data into broader categories. If the original scores cover a range of no more than, say, 20 points, grouping may not be called for.

In practice, when we are tabulating a set of data, deciding on the size of the score interval is the *first* step. Next we set up the score intervals, as shown in the left-hand column of Table 5.3. Each individual is then represented by a tally mark, as shown in the middle column. (It is easier to keep track of the tallies if every fifth tally is a diagonal line across the preceding four.) The column headed Frequency is gotten by counting the number of tallies in each score interval.

#### GRAPHIC REPRESENTATION

It is often helpful to translate the facts of Table 5.3 into a pictorial representation. A common type of graphic representation, which is called a *histogram*, is shown in Fig. 5.1. This can be thought of, somewhat grimly, as "piling up the bodies." The score intervals

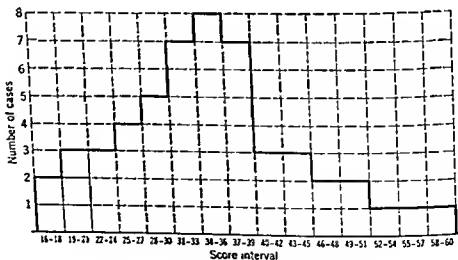


Fig. 5.1. Histogram of reading scores.

are shown along the horizontal base-line (abscissa). The vertical height of the pile (ordinate) represents the number of cases. The diagram indicates that there are two "bodies" piled up in the interval 16-18, three in the interval 19-21, and so forth. This figure gives a clear picture of how the cases pile up, with most of them in the 30's and a long low pile running up to the high scores.

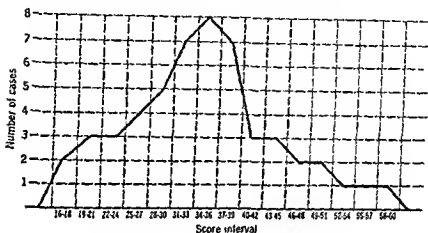


Fig. 5.2 Frequency polygon of reading scores.

Another way of picturing the same data is by preparing a *frequency polygon*. This is shown in Fig. 5.2. Here we have plotted a point at the mid-point of each of our score intervals. The height at which we have plotted the point corresponds to the number of cases, or frequency ( $f$ ), in the interval. These points have been connected, and the jagged line provides a somewhat different picture of the same set of data illustrated in Fig. 5.1. Histogram and frequency polygon are essentially interchangeable ways of showing the same facts.

## MEASURES OF CENTRAL TENDENCY

We often need a statistic to represent the typical, or average, or middle score of a group of scores. A very simple way of identifying the typical score is to pick out the score that occurs most frequently. This is called the *mode*. If we examine the array of scores on p. 99, we see that the score 36 occurs 4 times and is the mode for this set of data. We can also note another fact. The score values 38, 37, 32, 31, and 27 each occur 3 times. If there were 1 less 36 and 1 more 27, for example, the mode would shift by 9 points. The mode is

sensitive to such minor changes in the data and is therefore a crude and not very useful indicator of the typical score. In Table 5.3, where we have the grouped frequency distribution, the *modal interval* is the interval 34–36. This is as closely as we can identify the mode for data presented in this way.

### MEDIAN

A much more useful way of representing the typical or average score is to find the value on the score scale that separates the top half of the group from the bottom half. This is called the *median*. In our example, in which we have 52 cases, we want to separate the top 26 from the bottom 26 pupils. The required value can be estimated from the scores shown in Table 5.3. Starting with the lowest score, we count up until we have the necessary 26 cases. The “counting up” is best done in a systematic way, as shown in Table 5.4.

Table 5.4. Frequency Distribution and Cumulative Frequencies for Reading Scores

Score Interval	Frequency	Cumulative Frequency
58–60	1	52
55–57	1	51
52–54	1	50
49–51	2	49
46–48	2	47
43–45	3	45
40–42	3	42
37–39	7	39
34–36	8	32
31–33	7	24
28–30	5	17
25–27	4	12
22–24	3	8
19–21	3	5
16–18	2	2

Table 5.4 shows the cumulative frequencies as well as the frequency in each interval. Each entry in the column labeled Cumulative Frequency shows the total number having a score equal to or less than the highest score in that interval. That is, there are 5 cases scoring at or below 21, 8 scoring at or below 24, 12 scoring at or below 27, and so forth. As indicated, we wish to identify the point below which



50 per cent of the cases fall. Since 50 per cent of 52 = 26, we must identify the point below which 26 pupils fall.

We note that 24 individuals have scores of 33 or below. We need to include 2 more cases to obtain the required 26 cases. Note that in the next score interval (34-36) there are 8 individuals. We require only  $\frac{2}{8}$  or  $\frac{1}{4}$  of these individuals. Now how shall we think of these cases being spread out over the score interval 34-36? As we indicated on p. 101, a reasonable assumption is that they are spread out evenly over the interval. Then to include  $\frac{1}{4}$  of the scores, we would then have to go  $\frac{1}{4}$  of the way up from the bottom of the interval toward the top.

At this point we must define what we mean by a score of 34. In the first place, let us note that although test scores go by jumps of 1 unit, i.e., 34, 35, 36, we consider the underlying ability to have a continuous distribution taking all intermediate values. Thus, we do not get a score of 34.27, but this is only because our test does not register that precisely. Our definition will be that a score of 34 means closer to 34 than to either 33 or 35. That is, 34 will mean from  $33\frac{1}{2}$  to  $34\frac{1}{2}$ . This definition is somewhat arbitrary but is rather generally accepted in statistics textbooks. Our class interval 34-36 is really to be thought of as extending from  $33\frac{1}{2}$  to  $36\frac{1}{2}$ . Since we require  $\frac{1}{4}$  of the cases in this interval, we have  $\frac{1}{4}(36\frac{1}{2} - 33\frac{1}{2}) = \frac{1}{4} \times 3 = \frac{3}{4} = 0.75$ . We must add 0.75 to the value  $33\frac{1}{2}$ , which is the borderline between the 2 intervals. The median for this set of scores is  $33.5 + 0.75 = 34.25$ .

To compute the median, then,

1. Calculate the number of cases that represent 50 per cent of the total group. In our example 50 per cent of 52 is 26.
2. Accumulate the scores up through each score interval. The cumulative frequencies, as shown in Table 5.4, are 2, 5, 8, 12, 17, etc.
3. Find the interval for which the cumulative frequency is just less than the required number of cases. In our example the cumulation through the 31-33 interval is 24.
4. Find the score distance to be added to the top of this interval, in order to include the required number of cases, by the following operation:

$$\left( \frac{\text{Number of additional cases required}}{\text{Number of cases in next interval}} \right) \left( \frac{\text{Number of score points in interval}}{1} \right)$$

In our example this becomes  $(\frac{2}{8})(3) = 0.75$ .

5. Add this amount to the upper limit of the interval. We have for our data  $33.5 + 0.75 = 34.25$ . This score is the *median*, the score below which 50 per cent of the cases fall.

### PERCENTILES

The same procedure may be used to find the score below which any other percentage of the group falls. These values are all called *percentiles*. The median is the 50th percentile, i.e., the score below which 50 per cent of individuals fall. If we want to find the 25th percentile, we must find the score below which 25 per cent of the cases fall. Twenty-five per cent of 52 is 13. Thirteen cases take us through the interval 25-27, and include 1 of the 5 cases in the 28-30 interval. So the 25th percentile is computed to be  $27.5 + (\frac{1}{5})3 = 27.5 + 0.6 = 28.1$ . Other percentiles can be found in the same way. Percentiles have many uses, especially in connection with test norms and the interpretation of scores.

### ARITHMETIC MEAN

Another frequently used statistic for representing the middle of a group is the familiar "average" of everyday experience. Since the statistician speaks of all measures of central tendency as averages, he identifies this one as the *arithmetic mean*. This is simply the sum of a series of scores divided by the number of scores. Thus, the arithmetic mean of 4, 6, and 7 is

$$\frac{4 + 6 + 7}{3} = 5.67$$

In our example, we can add together the scores of all 52 individuals in our group. This gives us 1798. Dividing by 52, we get 34.58 for the "average" or arithmetic mean for this group.

Adding together all the scores and dividing by the number of cases is the straightforward way of computing the arithmetic mean. If the group is fairly small, and especially if an adding machine is available, it may be the best way. However, it can be rather laborious, especially with a large group. More efficient computing procedures are available, based on the frequency distribution given in Table 5.3. These calculations are based on a type of "trial balance." Picking a score interval that looks to be about in the middle of the group, we sum the plus and minus deviations from this starting place. An adjustment based on the excess of plus or minus deviations and applied

Table 5.5. Frequency Distribution of Reading Scores Showing Steps in Calculating Arithmetic Mean and Standard Deviation

Score Interval	Frequency $f$	$x'$	$fx'$	$f(x')^2$
58-60	1	8	8	64
55-57	1	7	7	49
52-54	1	6	6	36
49-51	2	5	10	50
46-48	2	4	8	32
43-45	3	3	9	27
40-42	3	2	6	12
37-39	7	1	7	7
34-36	8	0	0	0
31-33	7	-1	-7	7
28-30	5	-2	-10	20
25-27	4	-3	-12	36
22-24	3	-4	-12	48
19-21	3	-5	-15	75
16-18	2	-6	-12	72
			+61	
			-68	
Sum	52		-7	535

to this starting place gives the value for the mean. The application of this procedure to the reading test data is shown in Table 5.5, and the steps are outlined below.

1. Choose some interval for the arbitrary starting place or "origin." In this example the interval 34-36 has been chosen. Call this interval zero. (Note: Any interval can be chosen, and the final result will be the same. The particular interval chosen is purely a matter of convenience.)

2. Call the next higher interval +1, the one above that +2, etc.; call the next lower -1, the one below that -2, etc. These are shown in the column labeled  $x'$ . This column indicates the number of interval steps each interval is above or below our chosen starting point.

3. For each row, multiply the number of cases (frequency) by the number of steps ( $x'$ ) above or below the chosen origin. These products give the values in the column headed  $fx'$ . Note the minus signs in the lower half of the column. (Ignore the column headed  $f(x')^2$  for now. It refers to a later topic.)

4. Sum the values in the  $fx'$  column, taking account of the plus and minus signs. (Mistakes will be avoided if the plus entries are summed separately, the minus entries summed, and then the two part sums combined to give the final total.)

5. Sum the frequencies in the column headed Frequency (or  $f$ ) to give the total number of cases in the group. This is usually labeled  $N$ .

6. Divide the sum of the  $fx'$  values by  $N$ . Multiply by the number of score points in each interval. Add the result to the score corresponding to the mid-point of the zero interval. (Note that if the sum in step 4 is negative, adding it becomes in effect subtraction.)

These operations can be expressed by the following formula: \*

$$\text{Mean} = \left( \frac{\text{Sum of } fx'}{N} \right) (\text{Interval}) + \text{Arbitrary origin}$$

In our illustration the values become

$$\begin{aligned} \text{Mean} &= \left( \frac{-7}{52} \right) (3) + 35 \\ &= (-0.134)(3) + 35 \\ &= -0.40 + 35 \\ &= 34.60 \end{aligned}$$

Starting where we did, the minus deviations slightly overbalanced the plus ones. There was an excess of 7 on the minus side. Our starting point was a little too high. We had to shift it down  $\frac{7}{52}$  of 1 interval or  $\frac{7}{52} \times 3$  points of score to find a true balance point. Since the middle of our zero interval corresponded to a score of 35, we had to move down  $2\frac{1}{52}$  points below 35 to get the true balance point, the correct arithmetic mean.

The value 34.60 that we got in this way is almost the same as the 34.58 that resulted from adding all the scores together and dividing by the number of cases. The correspondence is usually not perfect, due to slight inaccuracies involved in grouping our scores into classes in the frequency distribution, but the values obtained by the two methods will always agree closely. It makes no difference which interval we use for our starting point. Barring mistakes in arithmetic, we will always get identically the same result.

The arithmetic mean and the median do not correspond exactly, but usually they will not differ greatly. In this example, the values are 34.60 and 34.25, respectively. The mean and median will differ substantially only when the set of scores is very "skewed," i.e., there is a piling up of scores at one end and a long tail at the other. Fig. 5.3 shows three distributions differing in amount and direction of

\* A list of common statistical symbols and their meanings is given at the end of the chapter. Reference to these definitions may help in reading the remainder of the chapter.

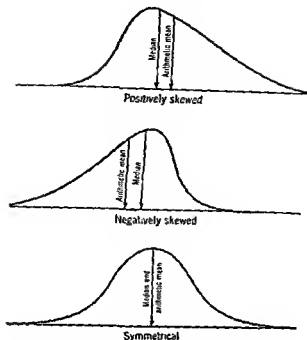


Fig. 5.3. Frequency distributions differing in skewness.

skewness. The top figure is positively skewed, i.e., has a tail running up into the high scores. We get a distribution like this for income in the United States, since there are many people with small and moderate incomes and only a few with very large incomes. The center figure is negatively skewed. A distribution like this would result if a class was given a very easy test, which resulted in a piling up of perfect and near-perfect scores. The bottom figure is symmetrical and is not skewed in either direction. Many physical and psychological variables give such a symmetrical distribution. In the many distributions that are approximately symmetrical either mean or median will serve equally well to represent the average of the group, but with skewed distributions the median generally seems preferable. It is less affected by a few cases out in the long tail.

## MEASURES OF VARIABILITY

When describing a set of scores, it is often significant to report how variable the scores are, how much they spread out from high to low scores. For example, two groups of children, both with a median age

of 10 years, would represent quite different educational situations if one had a spread of ages from 9 to 11 while the other ranged from 6 to 14. A measure of this spread is an important statistic for describing a group.

A very simple measure of variability is the *range* of scores in the group. This is simply the difference between the highest and the lowest score. In our reading test example it is  $59 - 17 = 42$ . However, the range depends only upon the 2 extreme cases in the total group. This makes it very undependable, since it can be changed a good bit by the addition or omission of a single extreme case.

#### SEMI-INTERQUARTILE RANGE

A better measure of variability is the range of scores that includes a specified part of the total group—usually the middle 50 per cent. The middle 50 per cent of the cases in a group are the cases lying between the 25th and 75th percentiles. We can compute these two percentiles, following the procedures outlined on pp. 105–106. For our example, the 25th percentile was computed to be 28.1. If we calculate the 75th percentile, we will find that it is 39.5. The distance between them is 11.4 points of score.

The 25th and 75th percentiles are called *quartiles*, since they cut off the bottom quarter and the top quarter of the group respectively. The score distance between them is called the *interquartile range*. A statistic that is often reported as a measure of variability is the *semi-interquartile range* ( $Q$ ). This is half of the interquartile range. It is the average distance from the median to the 2 quartiles, i.e., it tells how far the quartile points lie from the median, on the average. In our example, the semi-interquartile range is

$$Q = \frac{39.5 - 28.1}{2} = 5.7$$

If the scores spread out twice as far,  $Q$  would be twice as great; if they spread out only half as far,  $Q$  would be half as large. Two distributions that have the same mean, same total number of cases, and same general form, and that differ only in that one has a variability twice as large as the other are shown in Fig. 5.4.

#### STANDARD DEVIATION

The semi-interquartile range belongs to the same family of statistics as the median. Its computation is based upon percentiles. There are also measures of variability that belong to the family of the arithmetic

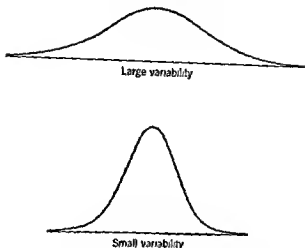


Fig. 5.4. Two distributions differing only in variability.

mean and are based upon score deviations. Suppose we had 4 scores which were 4, 5, 6, and 7 respectively. Adding these together and dividing by the number of scores we get

$$\frac{4 + 5 + 6 + 7}{4} = \frac{22}{4} = 5.5$$

This gives us the arithmetic mean. But now we ask how widely these scores spread out around that mean value. Suppose we find the difference between each score and the mean, i.e., we subtract 5.5 from each score. We then have  $-1.5$ ,  $-0.5$ ,  $0.5$ , and  $1.5$ . These represent *deviations* of the scores from the mean. The bigger the deviations, the more variable the set of scores. What we require is some type of average of these deviations to give us an over-all measure of variability.

If we simply sum the above 4 deviation values, we find that they add up to zero. This is necessarily so. We defined our arithmetic mean as the point around which the plus and minus deviations exactly balance. We shall have to do something else. The procedure that statisticians have devised for handling the plus and the minus signs is to square all the deviations. (A minus times a minus is a plus). An average of these squared deviations is obtained by summing them and dividing by the number of cases. To compensate for squaring the individual deviations, the square root of this average value is computed. The resulting statistic is called the *standard deviation* (SD

or  $s$ ). It is the square root \* of the average of the squared deviations from the mean. For our little example of 4 cases, the calculations are as follows:

$$\begin{aligned} SD &= \sqrt{\frac{(-1.5)^2 + (-0.5)^2 + (0.5)^2 + (1.5)^2}{4}} \\ &= \sqrt{\frac{2.25 + 0.25 + 0.25 + 2.25}{4}} = \sqrt{\frac{5}{4}} \\ &= \sqrt{1.25} = 1.12 \end{aligned}$$

#### STANDARD DEVIATION COMPUTED FROM FREQUENCY DISTRIBUTION

The standard deviation may also be computed from the grouped frequency distribution. The necessary steps have been carried out in Table 5.5. Take special note of the column headed  $f(x')^2$ . Each entry in this column represents the number of cases ( $f$ ) multiplied by the square of the deviation ( $x'$ ) of that score interval from the arbitrary origin. The sum of the values in this column gives a sum of squared deviations, but these deviations are around our arbitrary origin and are expressed in interval units. Several adjustments are necessary to express the deviations in *score* units and in terms of the *true* arithmetic mean. The steps are outlined below.

1. Carry out the operations for computing the arithmetic mean, as described on pp. 106-108.
2. In addition, prepare the column headed  $f(x')^2$ . Each entry in this column is the frequency ( $f$ ) times the square of the deviation value ( $x'$ ). However, this last column can be computed most simply by multiplying together the entries in the two preceding columns, i.e.,  $x'$  times  $fx'$ . Note that all the signs in this column are positive, since a minus times a minus gives a plus.

	<i>In symbolism</i>	<i>Illustrative example</i>
3. Get the sum of the $f(x')^2$ column. ("The sum of" will be indicated by $\Sigma$ .)	$\Sigma f(x')^2$	535
4. Divide this sum by the number of cases.	$\frac{\Sigma f(x')^2}{N}$	$\frac{535}{52} = 10.288$
5. Divide the sum of the $fx'$ column by the number of cases.	$\frac{\Sigma fx'}{N}$	$\frac{-7}{52} = -0.135$

\* The steps for computing the square root are shown in Appendix I.



	<i>In symbolism</i>	<i>Illustrative example</i>
6. Square the value obtained in 5 above.	$\left(\frac{\sum f x'}{N}\right)^2$	$\left(\frac{-7}{52}\right)^2 = (-0.135)^2 = 0.018$
7. Subtract the value in 6 from that in 4.	$\frac{\sum f(x')^2}{N} - \left(\frac{\sum f x'}{N}\right)^2$	$\frac{535}{52} - \left(\frac{-7}{52}\right)^2 = 10.288 - 0.018 = 10.270$
8. Take the square root of the value in 7.	$\sqrt{\frac{\sum f(x')^2}{N} - \left(\frac{\sum f x'}{N}\right)^2}$	$\sqrt{10.270} = 3.20$
9. Multiply by the number of score points in each class interval. (We call this width of interval 1.)	$1 \sqrt{\frac{\sum f(x')^2}{N} - \left(\frac{\sum f x'}{N}\right)^2}$	$3(3.20) = 9.60$

Presenting all the computations for our example in summary form, using the formula given in step 9 above, we have

$$SD = 3 \sqrt{\frac{535}{52} - \left(\frac{-7}{52}\right)^2} = 9.60$$

### INTERPRETING THE STANDARD DEVIATION

It is almost impossible to say in any simple terms what the standard deviation *is* or what it corresponds to in pictorial or geometric terms. Primarily, it is a statistic that characterizes a distribution of scores. It increases in direct proportion as the scores spread out more widely. The larger the standard deviation, the wider the spread of scores. A student sometimes asks: But what is a small standard deviation? What is a large one? There is really no answer to this question. Suppose that for some group the standard deviation of weights is 10. Is this large or small? It depends on whether we are talking about ounces, or pounds, or kilograms. It depends upon whether we are dealing with the weights of mice, or men, or mammoths. Large and small have only relative meaning—i.e., larger or smaller than that found for some other group or with some other test.

The standard deviation gets its most clear-cut meaning for one particular type of distribution of scores. This distribution is called the "normal" distribution. It is defined by a particular mathematical equation, but to the everyday user it is defined approximately by its pictorial qualities. The "normal" curve is a symmetrical curve having a bell-like shape. That is, most of the scores pile up in the middle score values; as one goes away from the middle in either direction the pile drops off, first slowly and then more rapidly, and the cases tail

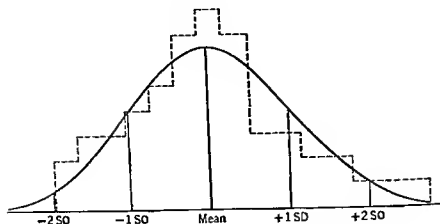


Fig. 5.5. Example of a normal curve (fitted to reading-test data).

out to relatively long tails on either end. An illustration of a typical normal curve is shown in Fig. 5.5. This curve is the normal curve that best fits the reading test data we have been using as an illustration. It has the same mean, standard deviation, and total area (number of cases) as the reading test data. The histogram of reading test scores appears in light dotted lines, so one can see how closely the curve fits the actual test scores.

For the normal curve, there is an exact mathematical relationship between the standard deviation and the proportion of cases. The same proportion of cases will always be found within the same standard deviation limits. This relationship is shown in Table 5.6. Thus, in *any* normal curve about two-thirds (68.2 per cent) of the cases will fall between  $+1$  and  $-1$  standard deviation from the mean. Approximately 95 per cent will fall between  $+2$  and  $-2$  standard

Table 5.6. Proportion of Cases Falling within Certain Specified Standard Deviation Limits for a Normal Distribution

Limits within Which Cases Lie	Per Cent of Cases
Between the mean and <i>either</i> $+1.0\ SD$ or $-1.0\ SD$	34.1
Between the mean and <i>either</i> $+2.0\ SD$ or $-2.0\ SD$	47.7
Between the mean and <i>either</i> $+3.0\ SD$ or $-3.0\ SD$	49.9
Between $+1.0$ and $-1.0\ SD$	68.2
Between $+2.0$ and $-2.0\ SD$	95.4
Between $+3.0$ and $-3.0\ SD$	99.8

deviations from the mean, and very nearly all the cases will fall between  $+3$  and  $-3$  standard deviations from the mean. An individual who gets a score 1 standard deviation above the mean will surpass 84 per cent of the group, i.e., he will surpass the 50 per cent who fall below the mean and the 34 per cent who fall between the mean and  $+1$  standard deviation.

This unvarying relationship of the standard deviation unit to the arrangement of scores in the normal distribution gives the standard deviation a type of *standard* meaning. It becomes a yardstick in terms of which different groups may be compared or the status of a given individual may be evaluated. Although the relationship of the standard deviation unit to the score distribution does not hold *exactly* in distributions other than the normal distribution, frequently the distribution of test scores or other measures approaches the normal curve closely enough so that the standard deviation continues to have very nearly the same meaning.

The meaning of being a given number of standard deviations above or below the mean may be expressed in terms of the per cent of cases in the group whom the individual surpasses. A number of values for this relationship are given in Table 5.7. This table provides a basis for interpreting any particular score. Consider the set of reading test scores for which we computed the mean and standard deviation to be 34.6 and 9.6 respectively. Suppose a person had a score of 40. Since the mean of the group is 34.6, he falls  $40 - 34.6 = 5.4$  points above

Table 5.7. Per Cent of Group Falling below Selected Standard Deviation Values for Normal Curve

Standard Deviation Value	Per Cent Having Scores below This Value
+3.0	99.9
+2.5	99.4
+2.0	97.7
+1.5	93.3
+1.0	84.1
+0.5	69.1
0.0	50.0
-0.5	30.9
-1.0	15.9
-1.5	6.7
-2.0	2.3
-2.5	0.6
-3.0	0.1

the mean of the group. The 5.4 points by which he surpasses the mean is equal to  $5.4/9.6 = 0.56$  standard deviations. He is 0.56 standard deviations above the mean. We might expect him to surpass approximately 71 per cent of the cases in our group. (An actual count shows that this score is better than  $3\frac{1}{2}\%_{52} = 75$  per cent of the scores in our set of data.) A score expressed in standard deviation units has much the same meaning from one set of scores to another, and these units are directly comparable from one measure to another.

In summary, the statistics most used for describing the variability of a set of scores are the semi-interquartile range and the standard deviation. The semi-interquartile range is based upon percentiles, i.e., the 25th and 75th percentiles, and is commonly used when the median is being used as a measure of the middle of the group. The standard deviation is a measure of variability that goes with the arithmetic mean. It is useful in the field of tests and measurements primarily as providing a standard unit of measure having comparable meaning from one test to another.

### INTERPRETING THE SCORE OF AN INDIVIDUAL

The problems of interpreting the score for an individual will be treated more fully in Chapter 6, when we turn to test norms and units of measure. It will suffice now to indicate that the two sorts of measures we have just been considering, i.e., percentiles and standard deviation units, each give us a framework in which we can view the performance of a specific person. Thus, referring to the example we worked out, if a new boy in the class got a score of 40 on the reading test we could say either

- a. That he surpassed 75 per cent of the group, i.e., that he fell at the 75th percentile, or
- b. That he fell 0.56 standard deviations above the mean.

Either statement gives his score meaning in relation to his group; he is somewhat above average but not one of the best ones in the group. Since they are based on the same score, they are two ways of saying the same thing. Each has certain advantages, which we will examine more carefully in Chapter 6.

### MEASURES OF RELATIONSHIP

We look now for a statistic to express the relationship between two sets of scores. Thus, in our illustration we have a reading score and

an arithmetic score for each pupil. To what extent did those pupils who did well in arithmetic also do well on the reading test? In this case, we have two scores for each individual. We can picture these scores by a plot in two dimensions. This is shown in Fig. 5.6. The first person in our group, Carol A, had a score on the reading test of 32 and a score on the arithmetic test of 3. Her scores are represented by the *X* in Fig. 5.6, plotted at 32 on the vertical or reading scale and at 3 on the horizontal or arithmetic scale. There is a dot to represent each other child's scores.

If a child who does well in reading also does well in arithmetic, we will find his scores represented by a dot in the upper right hand part of our picture. A child who does poorly on both tests will fall at the lower left. Where good score on one test goes with poor score on the other, we will find the points falling in the other corners, i.e., upper left and lower right. Inspection of Fig. 5.6 will show some tendency for the scores to splatter out in the lower-left to upper-right direction, i.e., from low-low to high-high. But there are many exceptions. The relationship is far from perfect. It is a matter of degree. We need some type of statistical index to express this degree of relationship. As an index of this degree of relationship, a statistic known as the

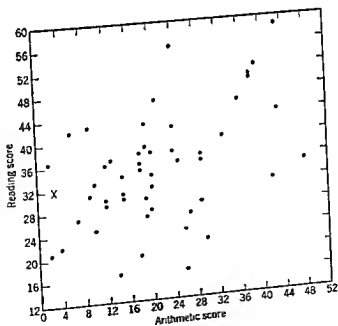


Fig. 5.6. Plot of reading versus arithmetic scores.

correlation coefficient can be computed. (The symbol  $r$  is used to designate this coefficient.) This coefficient can take values ranging from  $+1$  through zero to  $-1$ . A correlation of  $+1$  signifies that the person who had the highest score on one test also had the highest score on the other, the next highest on one was the next highest on the other, and so on, exactly in parallel through the whole group. A correlation of  $-1$  means that the scores go in exactly the reverse direction, i.e., the person highest on one is lowest on the other, next highest on one is next lowest on the other, etc. A zero correlation represents a complete lack of relationship. In-between values of  $r$  represent tendencies for relationship to exist but with many discrepancies.

Figure 5.7 illustrates four different levels of relationship. In box A the correlation is zero, and the points scatter out in a pattern that is just about round. All combinations are found—high-high, low-low, high-low, and low-high. Box B corresponds to a correlation of  $+.30$ .

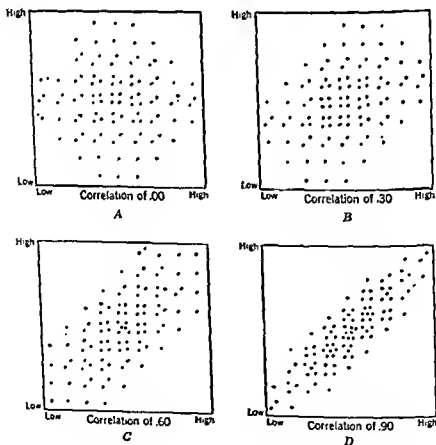


Fig. 5.7. Distribution of scores for representative values of correlation coefficient.

You can see a barely perceptible trend for the points to group in the low-low and high-high direction. The tendency is more marked in box C, which represents a correlation of  $+ .60$ . In box D, which portrays a correlation of  $+ .90$ , the trend is much more marked. But even with as high a correlation as this, the scores spread out quite a bit and do not follow an exact line from low-low to high-high. We may note in passing that the scores plotted in Fig. 5.6 correspond to a correlation coefficient of  $+ .46$ . Procedures for computing the correlation coefficient are outlined in Appendix II for those readers who wish to carry out the calculations with a numerical example.

There are two important settings in which correlation coefficients will be encountered in connection with tests and measurements. The first situation is one in which we are trying to determine how precise and consistent a measurement procedure is. Thus, if we wanted to know how consistent a measure of speed we get from a 50-yard dash, we could have each child run the distance twice, perhaps on successive days. The correlation of his two scores would give us information on the precision or reliability of this measure of running speed. The second situation is one in which we are studying the relationship between two different measures, often in order to evaluate one as a predictor of the other. Thus, we might want to study a scholastic aptitude test as a predictor of college grades. The correlation of test with grades would give an indication of the test's usefulness as a predictor.

We face the problem, in each case, of evaluating the correlation we obtain. Suppose the two sets of 50-yard dash scores yield a correlation of  $.80$ . Is this satisfactory or not? Suppose the aptitude test correlates  $.60$  with college grades. Shall we be pleased or discouraged?

The answer lies in part in the plots of Fig. 5.7. Clearly, the higher the correlation, the more closely one variable goes with the other. If we think of discrepancies away from the diagonal line from low-low to high-high as "errors," the errors become smaller as the correlation becomes larger. But, these discrepancies are still discouragingly large for even a rather substantial correlation coefficient, i.e., box C in Fig. 5.7. We must always be aware of these discrepancies and realize that with a correlation of  $.60$ , for example, between an aptitude test and school grades, there will be a number of children whose school performance differs a good deal from what we have predicted from the test.

However, everything is relative, and any given correlation coefficient must be interpreted in comparison to values that are commonly

obtained. Table 5.8 contains a number of different correlations that have been reported for different types of variables. The nature of the scores being correlated is described and the coefficient reported. An examination of this table will provide some initial background for interpreting correlation coefficients. The coefficient will gradually take on added meaning as the reader encounters coefficients of different sizes in his reading about and work with tests.

Table 5.8. Correlation Coefficients for Selected Variables

Variable	Correlation Coefficient
Height of identical twins	.95
Intelligence of identical twins	.88
Height versus weight	.60
Intelligence of siblings	.53
Height of siblings	.50
Strength of grip and speed of running	.16
Height versus Binet IQ	.06
Height versus educational achievement	.01
Shape of head versus intelligence	.01
Height versus sociability	.00
No. of physical defects among boys versus school progress	-.29

### SUMMARY STATEMENT

We opened this chapter by pointing out the various kinds of questions we might wish to answer by referring to a set of test scores. Let us look at these questions again and see what answers we have offered for them.

1. *How Do Our Scores "Run"; What Do They "Look Like"?* To answer this question, we can arrange our scores into a frequency distribution (Table 5.4) or plot them in a histogram (Fig. 5.1).
2. *What Score Is Typical of the Group; Represents the Middle of the Group?* To represent the middle of the group we may calculate the median—the 50th percentile (pp. 105–106), or the arithmetic mean—the common average (pp. 106–108).
3. *How Widely Spread Out Are the Scores; How Much Do They Scatter?* To represent the spread of scores statisticians have developed (1) the semi-interquartile range, half the distance between the 25th and 75th percentile (p. 110), and (2) the standard deviation



(pp. 111-113), a type of average of the deviations of the scores away from the average.

4. *How Are We to Determine What the Score of an Individual Means—Whether It Is High or Low?* Though this problem is left for fuller discussion in Chapter 6, we have seen that the individual score takes on meaning as it is translated into a percentile rank, the per cent of the group he beat, or into a standard score, how many standard deviations above or below the mean he fell (p. 116).

5. *To What Extent Do Two Sets of Scores Go Together; to What Extent Are the Same Individuals High or Low on Both?* A measure of relationship is given by the correlation coefficient, a numerical index of "going-togetherness" (pp. 116-120). This index is important as describing the precision or reliability of a test and as describing the accuracy with which a test score predicts some other factor, such as school grades or job success.

## STATISTICAL SYMBOLS

The student who reads test manuals, books dealing with tests, or articles about testing in the educational journals will encounter a number of conventional symbols to refer to statistical concepts or operations. Some of the commonest are defined below. This table of definitions should help in reading later chapters of this book, as well as outside references.

Symbol	Definition
$N$	The total number of cases in the group.
$f$	Frequency. The number of cases with a specific score or in a particular class interval.
$X$	A raw score on some measure.
$x$	A deviation score, indicating how far the individual falls above or below the mean of the group.
$x'$	A deviation score from some arbitrary reference point, often expressed in interval units.
$i$	The number of points of score in one class interval.
$\bar{X}$ or $M$	The mean of the group.
$Md$	The median of the group.
$Q_1$	The lower quartile, the 25th percentile.
$Q_3$	The upper quartile, the 75th percentile.
$Q$	The semi-interquartile range. Half the difference between $Q_3$ and $Q_1$ .
$P$	A percentile.
A subscript	Modifies a symbol and tells which specific individual or value is referred to, e.g., $P_{10}$ is the 10th percentile, $X_j$ is the raw score of person $j$ .

Symbol	Definition
SD or $s$	Standard deviation of a set of scores.
$\sigma$	Standard deviation in the <i>population</i> , though sometimes used to refer to the particular sample.
$p$	Per cent of persons getting a test item correct.
$q$	Per cent of persons getting a test item wrong ( $p + q = 100$ ).
$r$	A coefficient of correlation.
$r_{11}$	A reliability coefficient. The correlation between two equivalent test forms or two administrations of a test.
$\Sigma$	"Take the sum of."

### SUGGESTED ADDITIONAL READING

- Garrett, Henry E., *Elementary statistics*, New York, Longmans, Green, 1956.
- Guilford, J. P., *Fundamental statistics in psychology and education*, 3rd ed., New York, McGraw-Hill, 1956.
- Nelson, M. J., E. C. Denny, and A. P. Coladarci, *Statistics for teachers*, New York, Holt, 1956.
- Walker, Helen M., and Joseph Lev, *Elementary statistical methods*, rev. ed., New York, Holt, 1958.

### QUESTIONS FOR DISCUSSION

1. For each of the sets of scores indicated below, select what appears to you to be the most suitable class interval, and set up a form for tallying the scores:

Test	Na. of Cases	Range of Scores
Arithmetic	84	8 to 53
Reading Comprehension	57	15 to 75
Interest Inventory	563	68 to 224

2. In each of the following distributions, indicate (a) the size of the class interval, (b) the mid-point of the intervals shown, and (c) the real limits of the intervals (i.e., the dividing lines between them).

(1)	4-7	(2)	17-19	(3)	50-59
	8-11		20-22		60-69
	12-15		23-25		70-79
	.		.		.
	.		.		.
	.		.		.

3. Using the spelling scores given in Table 5.1 on p. 96, make a frequency distribution and a histogram. Compute the median and the upper and lower quartiles. Compute the arithmetic mean and standard deviation.

4. In the Bureau of Census reports the *median* is used in reporting average income. Why is it used, rather than the arithmetic mean?

5. A 50-item vocabulary test given to 150 pupils yielded scores ranging from 18 to 50. Ninety-seven fell between 40 and 50. What would this distribution of scores look like? What could you say about the suitability of the test for the group? What measure of central tendency would be most suitable? Why? What measure of variability would you probably use?

6. A high-school teacher gave two sections of a history class the same test. Results were as follows:

	<i>Section A</i>	<i>Section B</i>
Median	64.6	64.3
Mean	65.0	63.2
75th percentile	69.0	70.0
25th percentile	61.0	54.0
Standard deviation	6.0	10.5

From these data, what can you say about the two classes? What implications do the data have for teaching the two groups?

7. A test in social studies, given to 2500 tenth- and eleventh-grade students, had a mean of 52 and a standard deviation of 10.5. How many standard deviations above or below the mean would the following pupils fall?

Alice	48	Henry	60	John	31
Willard	56	Jane	36	Oscar	84

8. If the distribution in the previous example was approximately normal, about what per cent of the group would each of these pupils surpass?

9. Explain the meaning of each of the following correlation coefficients:

- The correlation between scores on a reading test and on a group intelligence test is  $+.78$ .
- Ratings of pupils on "good citizenship" and on "aggressiveness" show a correlation of  $-.56$ .
- The correlation between height and score on an achievement test is  $.02$ .

## Chapter 6



# Norms and Units for Measurement

### THE NATURE OF A SCORE

Johnny got a score of 15 on his spelling test. What does that mean, and how should we interpret it?

Actually, as it stands it has no meaning at all and is completely uninterpretable. At the most superficial level, we don't even know whether this represents a perfect score, i.e., 15 out of 15, or a very low per cent of the possible, i.e., 15 out of 50. But even supposing we do know that it is 15 out of 20, or 75 per cent, what then?

Look at Table 6.1. This shows two 20-word spelling tests. A score of 15 would have vastly different meaning if it were on test A than on test B. A person who got only 15 right on test A would not be outstanding in a second- or third-grade class. Try test B out on some friends or classmates. You will probably not find many of them who can spell 15 of these words correctly. When this test was given to a class of graduate students, only 22 per cent of them spelled 15 of the words correctly. A score of 15 on test B is a good score among graduate students of education.

As it stands, then, a score of 15 words right, or even of 75 per cent of the words right, can have no meaning or significance. It gets meaning only as we have some standard with which to compare it.

In the usual classroom test, the standard operates indirectly and imperfectly, partly through the teacher's choice of tasks to make up the test and partly through his standards for evaluating the responses. Thus, the teacher picks tasks to make up the test that he considers to be appropriate to represent the learnings of his group. No teacher in his right mind would give test A to a high-school group or test B to third graders. Where the responses vary in quality, as in essay examinations, the teacher sets a standard for grading that corresponds to what he considers it reasonable to expect from a group like his.

Table 6.1. Two 20-Word Spelling Tests

Test A	Test B
bar	baroque
cat	catarrh
form	formaldehyde
jar	jardiniere
nap	naphtha
dish	discernible
fat	fatiguing
sack	sacrilegious
rich	ricochet
sit	citrus
feet	feasible
act	accommodation
rate	inaugurate
inch	insignia
rent	deterrent
lip	eucalyptus
air	questionnaire
rim	rhythm
must	ignoramus
red	accrued

Quite different answers to the question "What were the causes of the War of 1812?" would be expected from a ninth grader and from a college history major.

However, the inner standard of the individual teacher is very subjective, inaccurate, and unstable. Furthermore, it provides no basis for comparing different classes or different areas of ability. We have no answers to such questions as: Are the children in school A better in reading than those in school B? Is Mary better in reading than in arithmetic? Is Johnny doing as well in algebra as we should expect? We need some broader, more uniform, objective and stable standard of reference if we are to interpret psychological and educational measurements.

Let us take a look at our tests A and B from another angle. Suppose, now, that we were to combine them into a single 40-word test and to give that test to 20 pupils in each grade from second through twelfth. What would we find? We would soon see that above the second or third grade almost everybody would get the first 20 words right. But until we got well up the grade ladder, children would get very few of the second set. It doesn't take much gain in spelling ability to improve from a score of 10 to one of 20 on this particular test, but

to improve from 20 up to a score of 30 represents quite a respectable accomplishment. The two 10-point gains don't begin to be equal. The units on our scale of scores cannot be considered equal units, then. We have a rubber yardstick that has been stretched out at some points and squeezed in at others.

There is one further point that we should make about our spelling scores. Let us consider test B, since the point will be most clearly and obviously true in this case. A person who fails to get any of the items right on test B cannot be said to fall at an absolute zero of spelling ability. Actually, he may be able to spell hundreds, possibly thousands, of words. So a person who gets 10 words right on test B doesn't demonstrate twice as much spelling ability as a person who gets only 5 right. On this test, as in an iceberg, the great bulk of what we are examining lies below "sea level" and can't be seen. We cannot guarantee that even test A gets down to a true zero point. In fact, it would be hard to say what a real zero point is in spelling ability.

### THE NEED FOR NORMS

We must look, then, for some better type of unit in which to express test results than a raw count of units of score or a crude percentage of the possible score. We would like the units to have these properties:

1. Uniform meaning from test to test, so that a basis of comparison is provided through which we may compare different tests—e.g., different reading tests, a reading test with an arithmetic test, or an achievement test with a scholastic aptitude test.
2. Units of uniform size, so that a gain of 10 points on one part of the scale signifies the same thing as a gain of 10 points on any other part of the scale.
3. A true zero point of "just none of" the quality in question, so that we can legitimately think of "twice as much as" or "two-thirds as much as."

The different types of norms that have been developed for tests represent marked progress toward the first two of the above objectives. The third can probably never be reached for the traits with which psychological and educational measurement is concerned. We can put five 1-pound prints of butter on one side of a pair of scales, and they will balance the contents of a 5-pound bag of flour poured into the other. "No weight" is *truly* "no weight," and units of weight can be added together. But we don't have that type of zero point or that way of adding together in the case of educational and psychologi-

The average height can be determined in the same way for 9-year-olds, 10-year-olds, and each other age group. The values will fall on some such curve as that shown in Fig. 6.1. Points for the curve will ordinarily be computed only for full-year groups, but the curve is to be considered continuous. That is, we can estimate points in between the year groups by referring to the continuous curve. Thus, in Fig. 6.1 a height of 60 inches corresponds to (or is average for) the age 12, while 50 inches corresponds to about 7 years and 8 months.

We can refer any height measurement to this scale and find for what age it would be average. Each girl's height can be interpreted as being the average height for a girl of a particular age. Thus, the

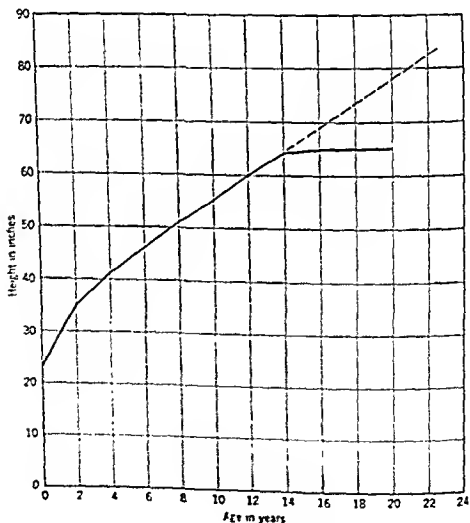


Fig. 6.1. Girls' age norms for height (Adapted from Boynton.)

girl who has a height of 60 inches can be described as being as tall as the average girl of 12 years. If we also know how old the girl actually is, we can judge whether she is tall, average, or short for her age. Thus, if Mary is 55 inches tall and is only 8 years old, we know that she is tall for her age. Her height is average for a 10-year-old.

The age framework is a relatively simple and familiar one. "He is as big as a 12-year-old" is a common way of describing a youngster. For a trait that shows continuous and relatively steady growth over a period of years, the age framework is a convenient one. Its familiarity and convenience are its major advantages. Age norms have a number of disadvantages, and these we must now consider in more detail.

The big issue in using age norms is whether we can reasonably think of a year's growth as representing a standard and uniform unit. Is the growth from age 5 to age 6 equal to the growth from age 10 to age 11, and similarly for each age on our scale? As we push up the age scale, we soon reach a point where we see that the year's growth unit is clearly inappropriate. There comes a point, some time in the teens or early 20's, when growth in almost any trait that we can measure slows down and finally stops. In Fig. 6.1 the slowdown takes place quite abruptly after age 14. A year's growth after 14 seems clearly to be much less than a year's growth earlier on the scale. After about 14 or 15, the concept of height-age ceases to have any meaning. The same problem of a flattening growth curve is found, varying only in the age at which it occurs and in abruptness, for any trait that we can measure. The failure of the unit "one year's growth" to have uniform meaning is most apparent as one considers the extremes of age, but there is no guarantee that this unit has uniform meaning even in the intermediate range.

The problem introduced by the flattening growth curve is most apparent when we consider the individual who falls far above the average. What is the height-age of a girl who is 5 feet 10 (70 inches)? The average woman *never* gets that tall at any age. If we are to assign any age value, we must invent some hypothetical extension of our growth curve such as the lightly dotted line in Fig. 6.1. This line assumes that growth continues at about the same rate that was typical up to age 14. On this extrapolated curve, the height of 5 feet 10 is assigned a height-age of about 16 years, 6 months. But this is a completely artificial and arbitrary age equivalent. It does *not* correspond to the average height of 16½-year-olds. It does not correspond to average height at any age. It merely signifies "taller than average."

This same type of artificial age equivalent must be used for ability



or achievement tests to express the performance of bright pupils in their teens. Mental growth curves also show a leveling off similar to that illustrated in Fig. 6.1. After the age of 14 or 15, increases become smaller and gradually disappear. The increase from age 15 to 18 may be no more than that from age 11 to 12, and after 18 there may be little or no further rise. Thus, a mental age of 18 or 20 does *not* mean performance corresponding to that of an average 18-year-old or 20-year-old. It is an arbitrary extension of the earlier growth curve. Such arbitrary and artificial age values are required if we are to be able to describe the performance of the upper half of our teen-age and adult population.

It is also true that growth curves are not entirely comparable for different functions. Rate of growth and time of reaching a maximum differ substantially. How shall we compare age scores on a vocabulary test and a maze-tracing test, for example, if the first continues to rise up to and into the twenties, while the second reaches a maximum in the early teens? For a 10-year-old to have reached the 12-year-old level may represent appreciably different degrees of superiority for different traits.

Two years' acceleration may also have quite different meaning, depending on the age level at which it occurs. A 5-year-old who is as tall as the 7-year norm is much more outstanding than the 10-year-old who reaches the 12-year norm. This fact has led to the development of the intelligence quotient and other types of quotients (which we shall consider presently) to allow for differences in age of the examinees. But the basic difficulty of inequality of the age unit at different points in the age scale still remains.

Of course, age norms are primarily appropriate for traits that depend on general normal growth. A trait showing no continuous improvement over an age range (such as acuity of vision) cannot possibly be expressed in terms of a scale of age units. One that depends primarily upon specific educational experiences, such as facility in arithmetical operations, seems to be more reasonably related to the educational framework of school grades than to the biological framework of years of growth.

Finally, though it does not directly concern the consumer of tests, it is worth noting that from the viewpoint of the test producer age norms present some serious practical problems. It is often difficult to get together a truly representative sample of individuals of a given age. Thus, if one wanted a cross-section of 12-year-olds one would have to look for some of them in the elementary school and some in the junior high school. They would have to be assembled from quite

a range of school grades. Then as one moves toward the older ages the sample one needs to reach is widely scattered—some in school, some at college, some in the military establishment, and some in the world of work. To reach a representative sample of 18-year-olds, for example, is a very forbidding task. This is one more reason why the usual age norms for tests become suspect as one moves up into the teens.

In summary, age norms, which are based on the performance of the average person at each age level, provide a readily comprehended framework for interpreting the performance of a particular individual. However, the equality of the age units is open to serious question. As one goes up to adolescence and adulthood, age ceases to have any meaning as a unit in terms of which to express level of performance. Age norms are most appropriate for the elementary-school years and for abilities that grow as a part of the general development of the individual. Physical and physiological characteristics such as height, weight, and dentition, and psychological traits such as general intelligence appear to be ones for which this type of norm is most acceptable.

## GRADE NORMS

Grade norms have many of the characteristics of age norms, differing only in that the reference groups are grade groups instead of age groups. That is, a test is given to representative groups in each of a series of school grades, and the average score is determined for each grade. Scores lying between the norm for two successive grades are assigned fractional credits by interpolation. The standard terminology assigns the value 5.0 to average performance at the beginning of the fifth grade, 5.5 to average performance at the middle of the grade, and so forth. A representative table of grade norms for the reading test of the *Metropolitan Achievement Test Battery* is shown in Table 6.3, p. 132. Thus, in this table a raw score of 9 corresponds to the performance of the average child at the beginning of the third grade, a raw score of 15 is average for beginning fourth grade, while 12 is average for the middle of grade three.

Grade norms have somewhat the same limitations as age norms. In particular, we have no guarantee that growth of one grade is the same amount of growth at all grade levels. The equality is even more suspect in the case of grade norms, because educational gains depend upon the content and emphasis in school instruction. The use of grade units to express growth only makes sense for those subject areas in which instruction is continuous through the school program. Since

Table 6.3. Grade Equivalents of Row Scores for Reading Test of Metropolitan Achievement Tests—Intermediate Battery, Form A

Raw Score	Grade Equiv.	Raw Score	Grade Equiv.
44	12.5	22	5.3
43	12.2	21	5.1
42	11.8	20	4.9
41	11.6	19	4.7
40	11.2	18	4.5
39	10.8	17	4.4
38	10.3	16	4.2
37	9.7	15	4.0
36	9.2	14	3.8
35	8.7	13	3.7
34	8.4	12	3.5
33	8.0	11	3.3
32	7.7	10	3.1
31	7.3	9	3.0
30	7.1	8	2.8
29	6.8	7	2.6
28	6.6	6	2.5
27	6.3	5	2.3
26	6.1	4	2.0
25	5.9	3	1.8
24	5.7	2	1.6
23	5.5	1	...

Reproduced by permission of Harcourt, Brace and World, Inc.

instruction in most of the basic skill subjects tapers off during high school, grade norms above the eighth or ninth have little direct meaning. In most cases, these are extrapolated values similar to those for the upper ages of age norms. Of course, grade norms for most high-school subjects would be essentially meaningless, since these are taught in only one or two grades.

The slowing down of gains at the upper grade levels makes it very difficult to express the performance of a very able child in terms of the grade framework. Many a superior child in the seventh or eighth grade can only be designated  $11+$  in terms of grade norms for standard school subjects. That is, his performance surpasses that of the average child in the highest grade for which norms are meaningful.

A further caution must be introduced with respect to the interpretation of grade norms. Consider a bright and educationally advanced child in the third grade. Suppose we find that on a standardized arith-

metic test he gets a score for which the grade equivalent is 5.9. This does *not* mean that our child has a mastery of the arithmetic taught in the fifth grade. He got a *score* as high as that gotten by the average child at the end of the fifth grade, but this higher score was almost certainly obtained in part by superior mastery of third-grade work. The average child is sufficiently slow and inaccurate that a number of score points (and consequently a higher grade equivalent) can be earned merely by real mastery at his own grade level. This is worth remembering. The fact that our child has a grade equivalent of 5.9 need not mean that the child is ready to move ahead into sixth grade work. It is only the reflection of a score and does not tell in what way that score was attained.

Grade norms are relatively easy to determine, since they are based on the administrative groups already established in the school organization. In the directly academic areas of achievement, the concept of grade level is perhaps a more meaningful one than age level. It is in relation to his grade placement that a child's performance in these areas is likely to be used and interpreted. Outside of the school setting, grade norms have little meaning.

To summarize, grade norms, which relate the performance of an individual to that of the average child at each grade level, are useful primarily in providing a framework for interpreting the academic accomplishment of children in the elementary school. For this purpose they are relatively convenient and meaningful, even though we cannot place great confidence in the equality of grade units. They have little value for other types of groups or measures.

## PERCENTILE NORMS

We have just seen that in the case of age and grade norms we give meaning to an individual's score by determining the age or grade group in which he would be just average. But it will often make more sense to compare him to his own age or grade group—to a group of which he may legitimately be considered a member. This is the type of comparison we make when we use percentile norms.

We saw in Chapter 5 how we could compute for any set of scores the median, quartiles, and any percentile. For each score value, we can compute the per cent of cases,  $p$ , falling below that score. Any person getting that score then surpasses  $p$  per cent of the group on which the percentile values were computed. We will say that he falls at the  $p$ th percentile, or has a percentile rank of  $p$ .

Table 6.4 shows percentile norms of ninth-grade boys for the eight

subtests of the *Differential Aptitude Test Battery*. Look at the column headed "Verb. Reas." (Verbal Reasoning). The entries are scores. Thus, a score of 24 corresponds to the 75th percentile. An individual who gets this score surpasses 75 per cent of the group on which the norms were based. A score of 17 corresponds to the 50th percentile on this test. On the *Abstract Reasoning Test* (Abs. Reas.), a score of 26 corresponds to the 50th percentile. This score represents the same degree of excellence as the score of 17 on the *Verbal Reasoning Test*.

Note that not every percentile is given in Table 6.4. For most of the range, the percentiles are given by steps of 5, and sometimes several score points correspond to the particular percentile value. If more detailed tables were given, these scores would correspond to different percentiles. However, locating an individual to the nearest 5 percentiles is close enough for all practical purposes.

Table 6.4. Percentile Norms for Differential Aptitude Tests

FORM A GRADE 9	BOYS		Raw Scores							N = 6000 ±	
	Percentile	Verb. Reas.	Num. Abl.	Abs. Reas.	Space Rel.	Mech. Reas.	Clerical SandA	LU-I: Spell.	LU-II: Sent.	Percentile	
	99	41+	35+	41+	87+	60+	73+	90+	59+	99	
	97	36-40	32-34	41-43	81-86	56-59	66-72	80-89	52-58	97	
	95	33-35	30-31	39-40	75-80	53-55	62-65	72-79	47-51	95	
	90	30-32	27-29	37-38	69-74	50-52	59-61	63-71	42-46	90	
	85	27-29	25-26	35-36	64-68	45-49	57-58	56-62	38-41	85	
	80	25-26	23-24	34	60-63	46-47	55-56	51-55	35-37	80	
	75	24	22	32-33	56-57	44-45	53-54	47-50	33-34	75	
	70	22-23	21	31	53-55	42-43	52	42-46	31-32	70	
	65	21	19-20	30	49-52	41	51	38-41	29-30	65	
	60	19-20	18	29	45-49	39-40	50	34-37	27-28	60	
	55	18	17	27-28	41-44	37-38	48-49	31-33	25-26	55	
	50	17	16	26	37-40	35-36	47	26-30	22-24	50	
	45	16	15	24-25	33-36	34	46	23-25	20-21	45	
	40	15	14	23	29-32	32-33	44-45	20-22	18-19	40	
	35	14	12-13	21-22	25-28	30-31	43	16-19	16-17	35	
	30	13	11	19-20	21-24	28-29	42	13-15	14-15	30	
	25	12	10	16-18	17-20	26-27	40-41	9-12	12-13	25	
	20	10-11	9	13-15	14-16	23-25	38-39	6-8	9-11	20	
	15	9	7-8	9-12	11-13	20-22	35-37	2-5	6-8	15	
	10	7-8	5-6	4-8	7-10	16-19	33-35	1	2-5	10	
	5	6	3-4	1-3	3-6	11-15	30-32	—	1	5	
	3	4-5	1-2	0	1-2	7-10	26-29	0	0	3	
	1	0-3	0	—	0	0-6	0-25	—	—	1	
	Mean	18.3	16.3	24.1	39.1	34.9	47.0	31.1	23.7	Mean	
	SD	8.7	8.2	11.3	23.4	12.6	10.5	24.1	14.6	SD	

Percentile norms are very widely adaptable and applicable. They can be used wherever an appropriate normative group can be obtained to serve as a yardstick. They are appropriate for young or old, for educational or industrial situations. To surpass 90 per cent of the reference comparison group signifies a comparable degree of excellence whether the function being measured is how rapidly one can solve simultaneous equations or how far one can spit. Percentile norms are widely used. Were it not for the two points that we must now consider, they would provide a very nearly ideal framework for interpreting test scores.

The first problem that faces us in the case of percentile norms is that of the norming group. On what type of group should the norms be based? Clearly, we will need different norm groups for different ages and grades in our population. A 9-year-old must be evaluated in terms of 9-year-old norms; a sixth grader, in terms of sixth-grade norms; an applicant for a job as stock clerk, in terms of stock-clerk-applicant norms. The appropriate norm group is in every case the group to which the individual belongs and in terms of which his status is to be evaluated. It makes no sense to compare a medical-school applicant with norms based on unselected adults.

If we are to use percentile norms, then, we must have multiple sets of norms. We must have norms appropriate for each distinct type of group or situation with which our test is to be used. This is recognized by the better test publishers, who provide norms not only for different age or grade groups but also for special types of educational or occupational populations. However, there are limits to the number of distinct groups for which a test publisher can produce norms.

Published percentile norms will often need to be supplemented by the test user, who can build up norm groups particularly suited to his individual needs. Thus, a given school system will often find it valuable to develop local percentile norms for its own pupils. This will permit interpretation of individual scores in terms of the local group, a comparison that may be more significant for local problems than comparison with the national norms. Again, an employer who uses a test with a particular category of job applicants may well find it useful to prepare norms for this particular group of people. Evaluating a new applicant will be much facilitated by these strictly local norms.

The second problem in relation to percentile norms is more serious. Again, we are faced by the problem of equality of units. Can we think of 5 percentile points as representing the same amount throughout the percentile scale? Is the difference between the 50th and 55th

percentile equivalent to the difference between the 90th and 95th? To answer this, we must notice the way in which test scores for a group of individuals usually pile up. We saw one histogram of scores in Chapter 5 (p. 102). This picture is fairly representative of the way the scores fall in many cases. There is a piling up of scores around the middle scores and a tailing off at either end. The ideal model of this type of score distribution, which is called the *normal curve*, was also considered in Chapter 5 (pp. 113-115) and is shown in Fig. 6.2. The exact normal curve is an idealized mathematical model, but many types of tests and measures distribute themselves in a manner that approximates a normal curve. You will notice the piling up of most of the cases in the middle, the tailing off at both ends, and the symmetrical pattern.

In Fig. 6.2, four score points have been marked. These are, in order, the 50th, 55th, 90th, and 95th percentiles. Note that near the median the 5 per cent of cases (the 5 per cent lying between the 50th and 55th percentile) fall in a tall narrow pile. Toward the tail of the distribution the 5 per cent of cases (the 5 per cent between the 90th and 95th percentile) make a relatively broad low bar. Five per cent of the cases spread out over a considerably wider range of scores in the second case than in the first. The same number of percentile points corresponds to about three times as many score points when we are around the 90th to 95th percentile as when we are near the median. The further out on the tail we go, the more extreme the situation becomes.

Thus, percentile units are typically and systematically unequal. The difference between being first or second in a group of 100 is many times as great as the difference between being 50th and 51st. Equal percentile differences do not represent equal differences in amount. Any interpretation of percentile ranks must take into account the fact

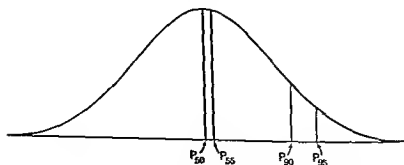


Fig. 6.2. Normal curve, showing selected percentile points.

that such a scale has been pulled out at both ends and squeezed in the middle. Mary, who falls at the 50th percentile in arithmetic and the 55th in reading, shows a trifling difference in these two abilities, whereas Alice, with percentiles of 90 and 95, shows a marked difference.

Percentile norms, to conclude, provide a basis for interpreting the score of an individual in terms of his standing in some particular group. If the percentile is to be meaningful, the group must be one with which it is reasonable and appropriate to compare him. We will usually need a number of tables of percentile norms based on different groups, if we are to use a test with different ages, grades, or occupations. As long as percentiles for appropriate groups are supplied, this type of norm is widely applicable. But interpretation of percentile values is made more difficult by the fact that we have a systematically "rubber" scale whose units are small in the middle range and large at the extremes.

## STANDARD SCORES

Because the units of a score system based on percentiles are so clearly not equal, we are led to look for some other unit that does have the same meaning throughout its whole range of values. *Standard-score* scales have been developed to serve this purpose.

In Chapter 5, we became acquainted with the standard deviation as a measure of the spread or scatter of a group of scores. The standard deviation was a type of average of the deviations of scores away from the mean—the root-mean-squared deviation. Scores may be expressed in standard deviations away from the mean. Thus, if the mean of a set of scores is 65 and the standard deviation is 15, a score of 80 is 1 standard deviation above the mean. A score of 35 is 2 standard deviations below the mean. In standard deviation units, we could call them  $+1.0$  and  $-2.0$  respectively.

Suppose we have given two tests to a group. The means and standard deviations for the group are shown below, as are the scores made by Johnny and Mary.

	Test A	Test B
Mean	65	40
Standard deviation	15	10
Johnny's score	77	55
Mary's score	87	48

Let us see how we can use standard scores to compare performances on the two tests or of the two individuals.



On test A, Johnny is 12 points above the mean, or  $12/15 = 0.8$  standard deviations above the mean. On test B he is 15 points, or  $15/10 = 1.5$  standard deviations above the mean. Thus, Johnny does a good deal better on test B than on test A. For Mary, the corresponding calculations give

$$\text{Test A: } \frac{87 - 65}{15} = 1.5 \qquad \text{Test B: } \frac{48 - 40}{10} = 0.8$$

Thus, we may say that Mary did as well on test A as Johnny did on test B, and vice versa. Each pupil's level of excellence is expressed as so many standard deviation units above or below the mean of the comparison group. This is a standard unit of measure having essentially the same meaning from one test to another. For aid in interpreting the degree of excellence represented by a standard score, see Table 5.7 (p. 115).

The type of score in standard deviation units that we have just presented is satisfactory except for two matters of convenience: (1) it requires us to use plus and minus signs which may be miscopied or overlooked, and (2) it gets us involved with decimal points which may be misplaced. We can get rid of the need to use decimal points by multiplying every standard deviation score by some constant, such as 10. We can get rid of minus signs by adding to every score a convenient constant amount such as 50. Thus, for Johnny's scores on test A and test B, we have

	Test A	Test B
Mean of distribution of scores	65	40
Standard deviation of distribution	15	10
Johnny's raw score	77	55
Johnny's score in standard deviation units	+0.8	+1.5
Standard deviation score $\times 10$	+8	+15
Plus a constant amount (50)	58	65

A table of standard scores for test A, based on this conversion, in which the mean is set equal to 50 and the standard deviation to 10, is shown in Table 6.5.

We could have used values other than 50 and 10 in setting up our conversion into convenient standard scores. The Army has used a standard-score scale with mean of 100 and standard deviation of 20 for reporting its test results. The College Entrance Examination Board has long used a scale with mean of 500 and standard deviation of 100. The Navy has used the 50 and 10 system.

Originally used in the Air Force, *stanine* scores have had some

Table 6.5. Standard-Score Equivalents for Test A

(Standard score mean = 50,  $SD = 10$ )

Raw Score	Standard Score	Raw Score	Standard Score	Raw Score	Standard Score
120	87	80	60	40	33
115	83	75	57	35	30
110	80	70	53	30	27
105	77	65	50	25	23
100	73	60	47	20	20
95	70	55	43	15	17
90	67	50	40	10	13
85	63	45	37	5	10

popularity in recent years. These are single-digit standard scores in which the mean is 5 and the standard deviation 2. The relationships among a number of the different standard score scales, and the relationship of each to percentiles and to the normal curve are shown in Fig. 6.3. The model of the normal curve is shown, and beneath it are a scale of percentiles and several of the common standard score scales. This figure illustrates the equivalence of scores in the different systems. Thus, a stanine score of 7 corresponds to an Army standard score of 120, a Navy standard score of 60, a College Board standard score of 600, a percentile rank of 84. The particular choice of score scale is arbitrary and a matter of convenience. It is too bad that all testing agencies have not been able to agree upon a common score unit. However, the important thing is that the same score scale and comparable norming groups be used for all tests in a given organization, so that results from different tests may be directly comparable.

Frequently standard-score scales are developed via the percentiles corresponding to the raw scores. The test maker assumes that the trait he is measuring is basically distributed in accordance with the normal curve. If he does not get a normal distribution of scores in his norming group, he assumes that this is because the raw-score units in which his test scores were expressed did not represent equal units throughout the range of scores. You will remember our discussion of this point in connection with our spelling test (pp. 125-126). He therefore takes steps to make his distribution of standard scores normal—he *normalizes* it. The actual calculations make use of percentiles and of tables of the normal curve. We shall not illustrate the details of procedure here.

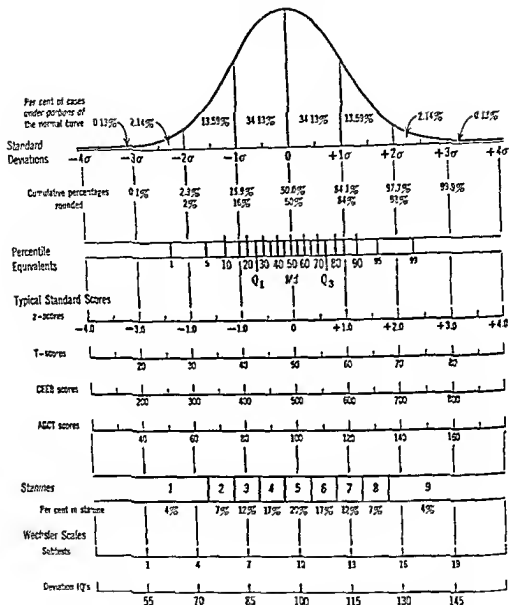


Fig. 6.3. Various types of standard-score scales in relation to percentiles and the normal curve. Reproduced by permission of the Psychological Corporation.

These standard scores have the distinctive feature that they are *guaranteed* to have a normal distribution, at least for a population comparable to that on which the original norms were obtained. The *score scale has been stretched in some places and squeezed together* in others so that finally a normal distribution results. This process of stretching and squeezing can take care of any inequality in the original units at different raw-score levels in the test. If the basic

assumption of a normal distribution was justified, this transformation will produce a scale in which a point of score really represents the same amount at any point on the scale. These are *normalized standard scores*. The term *T-score* which the reader of testing literature quite often encounters refers to this type of normalized standard score based on a single age group.

In summary, standard scores, like percentiles, base the interpretation of the individual's score on his performance in relation to a particular reference group. They differ from percentiles in that they are expressed in presumably equal units. The basic unit is the standard deviation of the reference group, and the individual's score is expressed as the number of standard deviation units above or below the mean of the group. Different numerical standard-score scales have been used by different testing agencies.

### INTERCHANGEABILITY OF DIFFERENT TYPES OF NORMS

Whichever type of norm is used, a table of norms will be prepared by the test publisher. This will show the different possible raw scores on the test, together with the corresponding score equivalents in the system of norms being used. Many publishers provide tables giving more than one type of score equivalent. An example is given in Table 6.6. Here we see the norms for the *Language Skills Test* of the revised *Metropolitan Achievement Test Battery*. Four types of norms are shown. The percentiles are based on a group tested early in the sixth grade. The standard-score scale assigns a mean of 50 and a standard deviation of 10 to a mid-sixth-grade group. Thus, a boy with a score of 21 can be characterized as

1. Having a grade equivalent of 8.6.
2. Falling at the 85th percentile in the sixth-grade group.
3. Receiving a standard score of 60.
4. Receiving a stanine of 7.

From Table 6.6 it is easy to see that the different systems of norms are different ways of expressing the same thing. We can translate from one to the other, moving back and forth. Thus, a child who falls at the 65th percentile in the sixth-grade group has a grade equivalent of 7.0. A grade equivalent of 7.0 corresponds to a standard score of 54. The different systems of interpretation support one another for different purposes.

Table 6.6. Norms for Language Study Skills Test of Metropolitan Achievement Tests—Intermediate Battery, Form A

Raw Score	Standard Score	Grade Equiv.	Grade 6 Percentile Rank (October testing)	Grade 6 Stanine (October testing)
28	80	12.5	—	
27	77	12.0	99½	
26	73	11.6	99	9
25	70	11.1	98	
24	67	10.6	95	
23	65	10.1	93	8
22	62	9.4	88	
21	60	8.6	85	7
20	58	8.0	80	
19	56	7.4	70	
18	54	7.0	65	6
17	52	6.6	60	
16	50	6.2	50	
15	49	5.9	45	5
14	47	5.6	40	
13	45	5.3	30	
12	43	5.0	25	4
11	42	4.8	20	
10	40	4.5	17	3
9	38	4.3	13	
8	36	4.0	10	
7	34	3.8	8	2
6	32	3.6	5	
5	30	3.3	2	
4	28	3.1	1	
3	24	2.9	1	1
2	20	2.6	1—	
1	16	2.4		

Reproduced by permission of Harcourt, Brace and World, Inc.

However, the different norm systems are not entirely consistent as we shift from one type of test to another. This is due to the fact that some functions mature more rapidly from one year to the next, relative to the spread of scores at a given age or grade level.

This can be seen most dramatically by comparing reading compre-

hension and arithmetic computation. The phenomenon is illustrated by the pairs of scores shown below that were taken from the Stanford

	Paragraph Meaning			Arithmetic Computation		
	John	Henry	Will	John	Henry	Will
Raw score	27	32	37	21	29	28
Grade equivalent	5.2	6.2	7.4	5.2	6.2	6.1
Grade 5.2 percentile	50	73	90	50	92	90

Achievement Test Battery. It is assumed that the three boys were tested at the end of 2 months in the fifth grade. John received scores on both tests that were just average. His grade equivalent was 5.2 and he was at the 50th percentile for pupils tested after 2 months in the fifth grade. Henry shows superior performance, but how does he compare in the two subjects? From one point of view, he does equally well in both; he is just one full year ahead of his grade placement. But in terms of percentiles he is much better in arithmetic than in reading, i.e., 92nd percentile as compared with 73rd percentile. Will, on the other hand, falls at just the same percentile for both reading and arithmetic. In his case, his grade equivalent for reading is 7.4 and for arithmetic is 6.1.

The discrepancies that appear in the above example are due to differences in the variability of performance and rate of growth of reading and arithmetic. Reading shows a *wide* spread within a single grade group, relative to the change from grade to grade. Some fifth graders read better than the average eighth or ninth grader, so a grade equivalent of 8 or 9 is not unheard of for fifth graders. In fact a grade equivalent of 8.0 corresponds to the 95th percentile for pupils at grade 5.2. By contrast, a fifth grader almost never does as well in arithmetic as an eighth or ninth grader—in part because he has not encountered or been taught many of the topics that will be presented in the fifth, sixth, seventh, and eighth grades. Thus, fifth graders are more homogeneous with respect to arithmetic skills, or looked at another way, arithmetic shows more rapid gains from fifth to eighth grade than does reading.

This point must always be borne in mind, especially in comparing grade equivalents for different subjects. A bright child will often appear most advanced in reading and language, least so in arithmetic and spelling—when the results are reported in grade equivalents. This difference may result, in whole or in part, simply from the differences

in the growth functions for the subjects, and need not mean a genuinely uneven pattern of progress for the child.

## QUOTIENTS

In the early days of mental testing, after age norms had been used for a few years, the need was felt to convert the age score into an index that would express rate of progress. The 8-year-old who had an age equivalent of  $10\frac{1}{2}$  years was obviously better than average, but how much better? Some index was needed to take account of chronological age (actual time lived) as well as the age equivalent on the test (score level reached).

The expedient was hit upon of dividing test age by chronological age to yield a quotient. This procedure was applied most extensively with tests of intelligence where the age equivalent we were concerned with was a mental age and the corresponding quotient was an *intelligence quotient*. However, it was also used to some extent for achievement tests and for some other sorts of measures.

The formula for computing the intelligence quotient in this way is given below and is illustrated for the 8-year-old who reaches the  $10\frac{1}{2}$ -year level on the test.

$$IQ = \frac{100MA}{CA}$$

$$= \frac{100(10.5)}{8} = 131$$

A similar quotient could be computed for a reading test, general achievement battery, measure of strength, or any other testing instrument that yields age norms. The resulting value would be called a reading quotient (RQ), educational quotient (EQ), or the like.

How does an intelligence quotient come to have meaning? In the first place, it is obvious by the way in which the quotient was established that 100 should be average at every age group, since the average 10-year-old, for example, should fall exactly at the 10-year level on any test if the age equivalents were properly established. But how outstandingly good is 125? How poor is 80? Such questions as these can only be answered by becoming acquainted with the distribution of quotients that a particular test yields.

The intelligence quotient was originally developed in connection with the individual intelligence test of the type represented by the

*Stanford-Binet* (see Chapter 9). A typical distribution of intelligence quotients for the 1937 revision of that test, based upon the standardization group, is shown in Table 6.7. This table shows the per cent of

Table 6.7. Distribution of Revised Stanford-Binet IQ's

IQ Range	Per Cent of Cases	Cumulative Per Cent
140 and over	1.3	99.9
130-139	3.1	98.6
120-129	8.2	95.5
110-119	18.1	87.3
100-109	23.5	69.2
90-99	23.0	45.7
80-89	14.5	22.7
70-79	5.6	8.2
60-69	2.0	2.6
Below 60	0.6	0.6

From L. M. Terman and M. A. Merrill, *Stanford-Binet intelligence scale*, Boston, Houghton Mifflin Co., 1960.

cases falling within each 10-point IQ interval and the cumulative percentage through each interval. Thus, 1.3 per cent of cases got IQ's of 140 and over, 3.1 per cent from 130 to 139, and so forth. An IQ of 125 would surpass roughly 91 per cent of the group (fall at the 91st percentile), whereas one of 80 would surpass only about 8 per cent. The mean for this particular distribution of IQ's is 101.5, and the standard deviation is 16.3.

The circumstance that made intelligence quotients from such a test as the *Stanford-Binet* relatively interpretable was that the mean and standard deviation remained relatively uniform from age to age. For this reason, an IQ of 125 signifies about the same status, relative to his own age group, whether obtained for a 5-year-old or a 15-year-old. This situation would not necessarily be true and was not perfectly true even for this test, but in many instances quotients were found to maintain the same average and spread of values in different age groups sufficiently closely so that a common interpretation was appropriate at all age levels.

To all intents and purposes, such quotients represent a type of standard score. In the case of the 1937 revision of the *Stanford-Binet*, we have a standard score with a mean of approximately 100 and standard deviation of approximately 16 in a general sample of American chil-



dren. This relationship of quotients to standard scores is explicitly recognized in most recent intelligence tests. For these, tables of IQ equivalents have been set up at each age level. These have been built so as to give a common mean and standard deviation for all age groups.

As a matter of fact, the most recent edition of the *Stanford-Binet*, brought out in 1960, also uses standard scores designed so that the mean is 100 and the standard deviation 16 at each age level, rather than the MA/CA ratio that was the basis for the IQ in earlier editions.

The quotients yielded by different tests are, unfortunately, not exactly equivalent. A variety of factors in the test and in the selection of norming groups have led to somewhat different means and standard deviations of intelligence quotients. Some evidence on the variability of quotients for five widely used tests for high-school groups is presented in Table 6.8. Experience with a test in a particular community

Table 6.8. Equivalent IQ's on Five Widely Used Group Intelligence Tests (From Engelhart<sup>2</sup>)

<i>Otis Quick- Scoring Beta, Form F<sub>m</sub></i>	<i>California Mental Ma- turity, Short Form, Inter- mediate Form S</i>	<i>Kuhlman- Anderson Battery Booklet G</i>	<i>Large- Thorndike Verbal, Level 4, Form A</i>	<i>Pintner General Ability, Intermediate Form A</i>
140	145	140	142	151
130	134	130	132	139
120	123	121	121	126
110	113	111	111	113
100	102	101	100	100
90	92	92	90	87
80	81	82	79	74
70	70	73	69	61

setting will provide a further basis for interpreting quotients at different levels.

The notion of the intelligence quotient or IQ is deeply imbedded in the history of the testing movement, and, in fact, in twentieth-century American culture. The expression "IQ test" is a part of our common speech. We are probably stuck with the term. But in the future IQ's will in most cases really be standard scores. And this is how we should think of them and use them. We may hope that eventually the test publishers will agree upon a common standard score scale and

will establish more clearly comparable normative groups, so that scores on different tests will be more directly comparable.

## PROFILES

The various types of norms we have been considering provide a means of expressing scores on quite different tests in common units in such a way that they can be directly compared. There is no direct way of comparing a score of thirty words correctly spelled with one of twenty arithmetic problems solved. But if both scores are expressed in terms of the grade level to which they correspond or in terms of the per cent of some defined common group that gets scores below that point, then they may be compared. The set of different test scores for an individual, expressed in a common unit of measure, constitute his *score profile*. The separate scores may be presented for comparison in tabular form by listing the converted score values. Illustrations of record forms showing the manner of recording converted scores are given in Figs. 6.4 and 6.5. The comparison of different subareas of performance is made pictorially clearer by a graphic profile. Several ways of plotting profiles are shown in Figs. 6.6, 6.7, and 6.8.

Figure 6.6 shows the form for plotting the subscores of the *California Test of Mental Ability*. Each subtest is represented by a row. The scale of age equivalents appears across the top of the form. The broken vertical line portrays the performance of the particular individual. Peaks in performance are to the right and low points to the left.

Figure 6.7 shows a similar form for plotting part scores on the *Metropolitan Achievement Test*. This form differs in representing the different tests in successive columns and presenting the score scale in the vertical dimension. Grade equivalents are shown on the vertical scale.

Figure 6.8 shows a type of profile chart for the component tests of the *Differential Aptitude Test Battery*. This battery undertakes to appraise different aspects of ability important in a high-school guidance program. Note that in this case the different tests are represented by separate bars, rather than points connected by a line. The scale used in this case is a percentile scale, but in plotting percentile values appropriate adjustments have been made for the inequality of percentile units. That is, percentile points have been spaced in the same way as they are in a normal curve, being more widely spaced at the upper and lower extremes than in the middle range. This percentile scale corresponds to the percentile scale that is shown in Figure 6.3 (p. 140). By this process, the percentile values for an individual are plotted on

## CLASS RECORD

In this Class Record achievement test results should be recorded in terms of grade or age equivalents. Mental Age and IQ's should be determined from an intelligence test.

In this Class Record achievement test results should be recorded in terms of grade or age equivalents. Minimum age 7 years.												
PUPIL'S NAMES	CHRONOLOGICAL AGE	MENTAL AGE	IQ	TESTS						AVE. ACHT		
				1. Read	2. Vocab.	Ave. Read	3. Arith. Probl.	4. Arith. Probl.	Ave. Arith.	5. Lang. Usage	6. Spelling	
1. Mary Anderson	11-8	12-3	105	6.5	6.7	6.6	5.6	6.0	5.8	6.5	6.3	6.3
2. Henry Baker	12-2	11-9	96	6.1	6.0	6.0	6.4	6.9	6.6	5.9	5.6	6.2
3. Carol Cohen	11-7	13-7	117	7.6	7.3	7.4	6.6	6.2	6.4	7.2	6.4	6.8
4. Harold Dominick	11-11	11-4	95	5.4	5.9	5.6	5.5	5.3	5.4	5.4	6.3	5.4
5. etc												
6												
7												
8												
9												
10												
11												
12												
13												
14												
15												

Fig. 6.4. Class record form for Metropolitan Achievement Test. (Scores recorded as grade equivalents.)

# California Test of Mental Maturity • Class Record Sheet



SERIES USED { Preprimary      Primary      Elementary      Intermediate      Advanced  
 Preprimary S.F.      Primary S.F.      Elementary S.F.      Intermediate S.F.      Advanced S.F.

School      Date Given      19      Teacher

City      Grade

PUPIL'S NAME	SEX	C. A.	M. A.			I. Q.			PERCENTILE RANK FOR AGE						
			TOTAL MENTAL	LANG-UAGE	NON LANG-UAGE	TOTAL MENTAL	LANG-UAGE	NON LANG-UAGE	MEMORY	SPATIAL RELATIONSHIP	LOGICAL REASONING	NUMERICAL REASONING	VERBAL CONCEPTS	TOTAL MENTAL	LANG-UAGE
1															
2															
3															
4															
5															
6															
7															
8															
9															
10															
11															
12															
13															
14															

Fig. 6.5 Form for recording Intelligence test results. (Reproduced by permission of California Test Bureau.)

... 4.4. Profile sheet for individual pupil. (Reproduced by permission of California Test Bureau.)

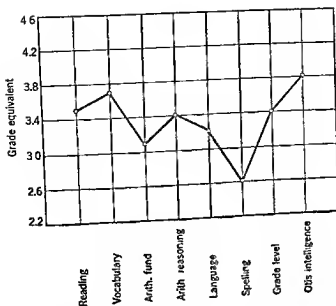


Fig. 6.7. Score profile for Metropolitan Achievement Test Battery and Otis Intelligence Test.

an equal-unit scale. A given linear distance can reasonably be thought of as representing the same amount of ability whether it lies high, low, or near the middle of the scale. By the same token, the same distance can be considered equivalent from one test to another.

Note that in Fig. 6.8 the bars have been plotted up and down from the 50th percentile. For this type of norm, the average of the group constitutes the anchor point of the scale, and individual scores can be referred to this base level. This type of figure brings out the individual's strengths and weaknesses very dramatically.

The profile chart makes a very effective way of representing the scores for an individual. In interpreting profiles, however, several cautions must be borne in mind. In the first place, procedures for plotting profiles assume that the norms for the several tests are comparable. Age, grade, or percentile scores must be based upon equivalent groups for all the tests. The best guarantee of equivalence is, of course, a common population used for all tests. This is the situation that commonly prevails for the different subtests of a test battery. Norms for all are established at the same time on the basis of testing a common group. The guarantee of comparability of the norms for the different component tests is one of the most attractive features of an integrated battery. If separately developed tests are plotted to-

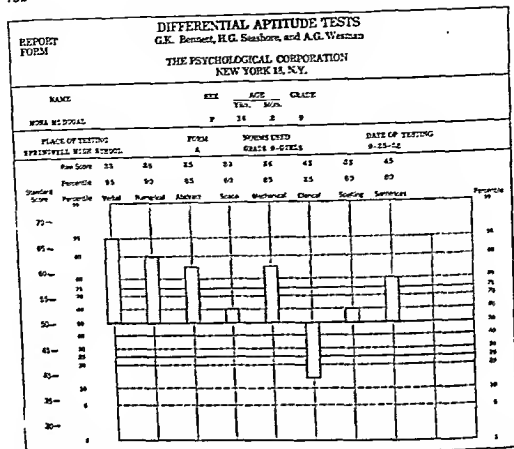


Fig. 6.8. Pupil profile chart for Differential Aptitude Tests. (Reproduced by permission of the Psychological Corporation.)

gether, we can usually only hope that the groups on which norms were established were comparable and that the profile is an unbiased picture of relative achievement in the different fields. Where it is necessary to use tests from several different sources, one solution is to develop our own local norms on a common population and to plot individual profiles in terms of the local norms.

A second problem is that of deciding how to interpret the ups and downs of a profile. Not all the differences that appear in a profile are meaningful, either in a statistical or a practical sense. We must decide which of the differences deserve some attention on our part and which should be ignored. This problem arises because no test score is completely exact. A full discussion of the problem of reliability and of the "error of measurement" in a test score will be provided in the following chapter. At this point, we shall merely note that test scores are not perfectly accurate, that performance on a reading test or an

aptitude test will vary somewhat from form to form and from occasion to occasion. Thus, small differences from one score to another in a test profile should be largely ignored as having very probably arisen by chance. Only as the differences between scores become substantial in relation to the standard error (see p. 175) of the separate scores is there any justification for interpreting the differences as representing something real and significant.

Organizing the separate test scores of an individual into a graphic profile is, then, a very effective way of dramatizing the high and low points in a score pattern. Such a profile may be plotted whenever scores from several different tests are expressed in the same units. However, a profile must be interpreted with a good deal of caution, because even unreliable differences may look quite impressive.

### USING NORMS

We have seen that norms provide a basis for interpreting the scores of an individual. Converting the score for any test taken singly into age or grade equivalent, percentile or standard score, permits an interpretation of the level at which the individual is functioning on that particular test. Bringing together the set of scores for an individual in a common unit of measure, and perhaps exhibiting these scores in a profile, brings out the relative level of performance of the individual in different areas.

The median performance for a class, a grade group in a school, or the children in a grade throughout a school system may be similarly reported. We then see the average level of performance within the group on some single function or the relative performance of the group in each of several areas. Norms provide a frame within which the picture may be viewed and bring all parts of the picture into the common frame. Now what does the picture mean, and what should we do about it?

Obviously we cannot, in a few pages, provide a ready-made interpretation for each set of scores that may be obtained in a practical testing situation. However, we can lay out a few general guiding lines and principles that may help to forestall some unwise interpretations of test results. The first points are phrased with an eye to the interpretation of group results. These are followed by some points relating primarily to interpretation of individual scores. However, the points overlap somewhat, and each has some reference to the other type of situation.



## PRINCIPLES GUIDING INTERPRETATION OF GROUP PERFORMANCE

1. *In Evaluating Average Group Achievement, Consideration Must Be Given To Average Ability Level in the Group.* A sixth-grade class with an average mental age of 10 years could not be expected to do arithmetic as well as one with an average mental age of 12 years. Some adjustment must be made for the typical ability level. However, one must be somewhat conservative in making such adjustments, especially for classes superior on an intelligence test. The correspondence between intelligence and academic achievement is not perfect, and a group of bright youngsters will rarely be comparably outstanding in achievement. This will be true particularly in the more specialized and less academic subjects, such as spelling or handwriting. A group that deviates from average in ability can be expected to differ from the general norm in achievement also, and in the same direction, but it should not be expected to differ as much in achievement as it does in ability.

2. *A Further Factor That May Be Expected To Influence Achievement Is the Type of Cultural Background from Which the Children Come.* Home and community influences are strong. Foreign-language background, absence of pictures and books in the home, a negative family attitude toward schools and schooling may all be important. In a measure, these factors affect intelligence test score. But they affect achievement also, and perhaps more directly. Where a class is atypical in cultural background, either especially favored or especially deprived, allowance must be made for this in interpreting test results.

3. *Group Achievement Can Only Be Evaluated in the Light of Curricular Content, Emphases, and Objectives.* If a school system has delayed all formal instruction in arithmetic until the third grade in order to provide more time in the earlier grades for group projects, social experiences, and preparatory materials, it is unreasonable to expect the children in the third grades of that system to come up to national third-grade norms in arithmetic. If a school system has de-emphasized accurate spelling as an objective, has cut down or eliminated spelling drills, and has concentrated on other educational outcomes, it is inappropriate to evaluate that school by rigid application of national norms in a standardized spelling test. There is a good deal of evidence from test results themselves that schools in the more prosperous and privileged communities have de-emphasized the basic tool skills of arithmetic and spelling in the early grades. In these grades such communities often do no better in computation and spelling than much poorer communities with children of lower intelligence.

Of course, the communities giving less emphasis to arithmetic and spelling in order to achieve other less tangible educational outcomes may not actually be achieving them. Whether they are can only be answered as we develop measures to appraise such objectives as ability to follow directions, to work alone, to take care of property, to get along with other children, or to grow in social relationships, which are objectives given emphasis in the stated objectives of these communities. Instruments for appraising these objectives should receive the attention of the measurement specialist and the schools themselves. But one thing is clear. The school's objectives and curricular emphases must be taken into account in interpreting standardized test results.

4. *Use of Test Results Should Be Constructive, Not Punitive.* One continually encounters situations in which results on achievement tests are used as a basis for evaluating the professional worth of teachers. The test then becomes a sword held over the teacher's head, a recurring threat to his security. In such a situation, it should be no wonder if the test is resented, if the teacher teaches in order to "beat the test" or even gives illicit help at the time of testing. The teacher is now on the side of the pupils working against the test.

This type of situation is to be avoided at all costs. The threat arises in large measure out of administrative personnel and will disappear if administrators see the tests as primarily tools to help both pupil and teacher. This will be facilitated if tests are given in the fall, when they can be used to guide the work of the year to come, rather than in the spring, to judge the work of the past year.

#### PRINCIPLES GUIDING INTERPRETATION OF INDIVIDUAL PERFORMANCE

1. *Here Again, Achievement Must Be Evaluated in the Light of Evidence of Aptitude.* The 12-year-old who is reading at the 9-year level is not a reading problem if his mental age is also 9. He is then doing about what could be expected of him. Too many remedial classes are filled with children who are really performing at or even above the level that should be expected for them.

Again, the intellectually superior child cannot generally be expected to be as superior in achievement as he is on the measure of intelligence. In the first place, achievement depends upon exposure. Even the bright fourth grader cannot be expected to do sixth-grade arithmetic if he has never encountered or been taught the processes. In some subjects, at least, opportunity sets very real limits to the level that a person can reach. In the second place, abilities are to a degree specialized. The child who is picked out because he is bright is likely to be somewhat less outstanding in other specialized educational skills.

2. *For Individuals as Well as Groups, We Must Take Account of Family and Cultural Differentials.* The wide range of variation in language background, richness of home resources, and incentives to progress in school may be expected to have a great impact on educational skills and accomplishments, and allowance for IQ differences will only in part take account of these factors.

3. *The Individual Child's Performance, Too, Must Be Judged in Terms of the Curriculum To Which He Has Been Exposed.* The individual pupils cannot be expected to progress as rapidly in those areas in which teaching emphasis is less. Furthermore, in those skills that are closely dependent upon instruction, even the able pupils cannot be expected to move ahead at a tempo much faster than that at which the material is presented. Thus, the bright child may be expected to be more advanced in word knowledge and reading skills, which he can readily pick up on his own, than in the processes of arithmetic, which he is unlikely to master until he has been exposed to them in the school setting.

4. *In the Case of the Single Individual, We Must Be Acutely Aware of the Existence of Errors of Measurement.* A test score does not identify the exact level of ability for the child. It represents the most likely value within a fairly broad band of possible values. Differences between areas of achievement must be viewed as tentative as long as these bands overlap. Differences between standing on two testings—say, two reading tests a few months apart—should not excite us unduly unless they are quite substantial. We should be rather conservative in “explaining” differences that may represent nothing more than the fallibility of our measuring instrument.

5. *In the Very Nature of Things, by the Way Test Norms Are Developed We Must in General Expect Half of a Group to Fall Below the Norm.* The norm is the average, the typical. It is neither the ideal of satisfactory accomplishment nor the standard to which we can hold everybody. It is the typical performance of typical individuals at the present time. In any average there must be as many below as above. Educators must avoid the compulsion to bring everybody “up to the norm.” We must be careful not to try to fit everybody into the Procrustean bed of the average.



## SUMMARY STATEMENT

A raw score, taken by itself, has no meaning. It gets meaning only by comparison with some reference group or groups. The comparison may be with:

1. A series of age groups (age norms).
2. A series of grade groups (grade norms).
3. A single group, indicating what per cent of that group the score surpassed (percentile norms).
4. A single group, indicating standard deviations above or below the group mean (standard scores)

Each alternative has certain advantages and certain limitations, which we have considered

To get an index of brightness from age norms, quotients such as the Intelligence quotient and educational quotient were devised. These become meaningful and usable when they have approximately the same standard deviation for all age groups. In that case, they are essentially standard scores and should be thought of as such.

If the norms available for a number of different tests are of the same kind and are based on comparable groups, all the tests can be expressed in comparable terms. They can then be shown pictorially in the form of a profile. Profiles emphasize score differences within the individual. When profiles are used, care must be taken not to over-interpret minor ups and downs of the profile.

Norms represent a descriptive framework for interpreting the score of an individual, a class group, or some large aggregation. However, before a judgment can be made as to whether an individual or group is doing well or poorly, allowance must be made for ability level, cultural background, and curricular emphases. The norm is merely an average, not a strait jacket into which all can be forced to fit.

## REFERENCES

1. Boynton, Bernice, The physical growth of girls, *Univ. Ia. Stud. Child Welf.*, 12, No. 4, 1936.
2. Engelhart, Max D., *Equivalence of intelligence quotients of five group intelligence tests*, Bureau of Pupil Guidance, Chicago Public Schools (mimeographed report, 10 pp., no date).

## SUGGESTED ADDITIONAL READING

- Flanagan, J. C., Units, scores, and norms, Chapter 17 in E. F. Lindquist, Editor, *Educational measurement*, Washington, D. C., American Council on Education, 1951.
- Harris, Chester W., Editor, *Encyclopedia of educational research*, 3rd ed., New York, Macmillan, 1960, pp. 922-926.
- Mosier, Charles L., Batteries and profiles, Chapter 18 in E. F. Lindquist,

Editor, *Educational measurement*, Washington, D. C., American Council on Education, 1951.

Seashore, Harold G., *Methods of expressing test scores. Test Service Bulletin No. 48*, New York, Psychological Corp., 1955.

## QUESTIONS FOR DISCUSSION

1. A pupil in the seventh grade received a raw score of 13 on the *Metropolitan Reading Test, Intermediate Level*. What additional information would be needed to interpret this score?

2. Why do standardized tests designed for use with high-school students almost never use age or grade norms?

3. What limitations would national norms have for use by a county school system in rural West Virginia? What might the local school system do about it?

4. What assumption or assumptions lie back of the development of age norms? Grade norms? Normalized standard scores?

5. In Fig. 6.8, p. 152, why are the standard scores evenly spaced whereas the percentile scores are unevenly spaced?

6. Using Tables 6.3 and 6.6, briefly characterize the following entering sixth-grade children:

	CA	MA	Reading Score	Study Skills
Pupil A	12.4	10.6	23	13
Pupil B	10.5	13.2	31	19
Pupil C	11.3	11.1	22	16

7. You are a guidance counselor and have given the *Differential Aptitude Battery* to a ninth grade. Using Table 6.4, prepare a summary report and interpretation for a boy with the following scores:

Verbal Reasoning	18	Mechanical Reasoning	54
Numerical Ability	23	Clerical Speed and Acc.	45
Abstract Reasoning	31	Spelling	14
Spatial Relations	72	Sentences	22

8. School A gives a battery of achievement tests each May in each grade from the third through the sixth. The median grade level in each subject in each teacher's class is reported to the superintendent. Should they be reported? If so, what else should be included in the report? In what ways might a superintendent use the results to advantage? What uses should he avoid?

9. Miss B prides herself that each year she has gotten at least 90 per cent of her fifth-grade group "up to the norm" in each subject. How desirable is this as an educational objective? What limitations or dangers do you see in it?

10. School C operates on a policy of assigning transfer students to a grade on the basis of their average grade standing on an achievement battery. Thus, a boy with a grade score of 6.4 on the battery as a whole would be assigned to the sixth grade, no matter what his age or his grade

in his previous school. What values do you see in this plan? What limitations?

11. The superintendent of schools in city D noted that school E fell consistently about a half grade below national norms on an achievement battery. He was distressed because this was the lowest of any school in his city. How justified is his dissatisfaction? What more do you need to know to answer this?

12. The board of education in city F noted that the second and third grades in their community fell substantially below national norms in arithmetic, though coming up to the norms in other subjects. They propose to study this further. What additional information do they need?

13. Look at the manual for some test, and study the information that is given about the norms.

- a. How adequate is the norming population? Is adequate information given about this?
- b. Figure out the chance score (i.e., the score to be expected from blind guessing) for each test, and note its grade equivalent. What limitations does this suggest on use of the test?
- c. What limitations are there on the usefulness of the test at the upper end of its range?
- d. How many raw score points correspond to one full grade?

14. Examine Fig. 6.6. What are the possible advantages of a profile such as this? What are its limitations and shortcomings? Is it desirable to plot it and use the results?

## Chapter 7



# Qualities Desired in Any Measurement Procedure

Whenever a worker in psychology or education desires to measure some quality in a group or individual, he faces the problem of choosing the best instrument for his purpose. Ordinarily there will be several tests or testing procedures that have been developed for, or that seem to be at least possibilities for, his purpose. He must choose among these. He is also probably interested in determining not only which is the best procedure but how well it satisfies his needs by some absolute standard. On what grounds can he make his choice or his appraisal?

There are many specific considerations entering into the evaluation of a test, but we shall consider them here under three main headings. These are respectively validity, reliability, and practicality. Validity refers to the extent to which a test measures what we actually wish to measure. Reliability has to do with accuracy and precision of a measurement procedure. Indices of reliability give an indication of the extent to which a particular measurement is consistent and reproducible. Practicality is concerned with a wide range of factors of economy, convenience, and interpretability that determine whether a test is practical for widespread use. These three aspects of test evaluation will be considered in detail in the following sections.

### VALIDITY

The first and foremost question to be asked with respect to any testing procedure is: How valid is it? When we ask this question, we are inquiring whether the test measures what we want it to measure, all of what we want it to measure, and nothing but what we want it to measure.

When we apply a steel tape measure to the top of our desk to determine its length, we have no doubt that the tape does in fact meas-

ure the length of the desk and does directly serve our purpose, which may be to determine whether the desk will fit between two windows in our room. Long experience with this type of measuring instrument has confirmed beyond a shadow of doubt its validity as a tool for measuring length.

Suppose now that we give to a group of children a test of reading achievement. This test requires the children to select certain answers to a series of questions about reading passages and to make little pencil marks on an answer sheet. We count the number of pencil marks made in the predetermined right places and give the child as a score the number of his right answers. We call this score his reading comprehension. But the score itself is not the comprehension. It is the record of a sample of behavior. Any judgment regarding comprehension is an inference from this number which is the number of allegedly correct answers. Its validity is not self-evident but is something we must establish on the basis of adequate evidence.

Consider again the typical personality inventory that endeavors to provide an appraisal of "emotional adjustment." In this type of inventory the respondent marks a series of statements as being characteristic of him or not characteristic of him. On the basis of various types of procedures, which we shall consider in some detail in Chapter 12, certain responses are keyed as indicative of emotional maladjustment. A score is obtained by seeing how many of these responses an individual selects. But making certain marks on a piece of paper is a number of steps removed from actually exhibiting emotional disturbance. We must find some way of establishing the extent to which the performance on the test actually corresponds to the quality of behavior in which we are directly interested. How can we determine the validity of such a measurement procedure?

#### TYPES OF EVIDENCE OF VALIDITY

A test may be thought of as corresponding to some aspect of human behavior in any one of three senses. For these three senses we shall use the terms (1) represent, (2) predict, and (3) signify. Let us explore each of these three, so that we may understand clearly what is involved in each case, and for what kinds of tests each of the three is relevant.

#### VALIDITY AS REPRESENTING

Consider a test that has been prepared to measure achievement in using the English language. How can we tell how well the test does in fact measure that achievement? First, we must reach some agreement



as to the skills, knowledge and understanding that comprise correct and effective use of English, and that have been the objectives of language instruction. Then we must examine the test to see what skills, knowledge and understanding it calls for. Finally, we must match the analysis of test content against the analysis of course content and instructional objectives and see how well the former *represents* the latter. In proportion as the outcomes that we have accepted as goals for the course are represented in the test, the test is valid.

Since the analysis is essentially a rational and judgmental one, this is sometimes spoken of as *rational or logical validity*. Since the analysis is largely in terms of the content of the test, the term *content validity* is also sometimes used. However, we should not think of content too narrowly, because we may be interested in *process* as much as in simple content. Thus, in the field of English expression we might be concerned on the one hand with such "content" elements as the rules and principles for capitalization, use of commas, or spelling words with "ei" and "ie" combinations. But we might also be interested in such "process" skills as arranging ideas in a logical order, writing sentences that present a single unified thought, or picking the most appropriate word to convey the desired meaning. In a sense, *content* is what the pupil works with; *process* is what he does with it.

The problem of appraising content validity is closely parallel to the problem of preparing the blueprint for a test, as discussed in Chapter 3, and then building a test to match the blueprint. A teacher's own test has content validity to the extent that a wise and thoughtful analysis of course objectives has been made in the blueprint, and care, skill and ingenuity have been exercised in building test items to match the blueprint. A standardized test may be shown to have validity for a particular school or a particular curriculum insofar as the tasks that it presents to the examinee correspond to and represent the objectives accepted in that school or that curriculum.

It should be clear that validity evidenced as *representing*, i.e., rational or content validity, is important primarily for measures of achievement. When we wish to appraise a test of reading comprehension, of biology, or of American history, we can really do so only by asking: How well do the tasks of this test represent what we consider to be important outcomes in this area of instruction? How well do these tasks represent what the best and most expert judgment would consider to be important knowledge and skills? If the correspondence is good, we consider the test valid; if poor, the validity must be deemed to be low.

The responsible maker of a test for publication and widespread use

goes to considerable pains to determine the widely accepted goals of instruction in the field in which his test is to be built. There are many types of sources to which he may, and often does resort. These include, among others: (1) the more widely used textbooks in the field, (2) recent courses of study for the large school units, i.e., states, counties, and city systems, (3) reports of special study groups, often appearing in yearbooks of one or another of the educational societies, (4) groups of teachers giving instruction in the course, (5) specialists in universities, cities, and state departments concerned with the training or supervision of teachers in the field.

Gathering information from these sources the test maker develops the blueprint for his test, and in terms of this blueprint he prepares his test items. Because of variations from community to community, no published test can be made to fit exactly the content or objectives of every local course of study. In this sense, a test developed on a national basis is always less valid for a specific community than an equally workmanlike test tailored specifically to the local situation. However, the well-made commercial test takes the common components that appear repeatedly in different textbooks and courses of study and builds a test around them. It *represents* the common core that is central in the different specific local patterns.

It should be clear from what has just been said that the relationship between teaching and testing is typically intimate. Test content is drawn from what has been taught, or what is proposed to be taught. The instructional program is the original source of test materials. Sometimes the thinking in a test may lead the thinking underlying a local course of study, as when a group of specialists have been brought together to design a test corresponding to some emerging trend in education. Sometimes the test may lag behind, as when the test is based on the relatively conventional objectives emphasized in established textbooks. But usually test content and classroom instruction are in close relationship to one another, and the test may be appraised by how faithfully it corresponds to the significant goals of instruction.

#### VALIDITY AS PREDICTING

Frequently we are interested in using a test to *predict* some specific future outcome. We use a scholastic aptitude test to predict how likely the high school student is to be successful in college X, where success is represented at least approximately by grade-point average. We use an employment test to pick machine operators who are likely to be successful employees, as represented by some such criterion as high production with little spoilage and low personnel turnover. For this

purpose, we care very little what a test looks like.\* We are interested only in the degree to which it correlates with some chosen criterion measure of job success. The higher the correlation, the better the test.

Our evaluation of a test as predicting is primarily an empirical and statistical evaluation, and this aspect of validity has sometimes been spoken of as *empirical* or *statistical validity*. The basic procedure is to give the test to a group who are entering some job or training program, to follow them up later and get for each one some criterion measure of success on the job or in the training program, and then to compute the correlation between test score and criterion measure of success. The higher the correlation, the more effective the test as a predictor.

This relationship can also be pictured in various ways. For example, the bar chart in Fig. 7.1 shows the percentage of persons failing pilot training at each of nine score levels on a predictor-test battery. Examination of the chart shows a steady increase in the per cent failing training as we go from the high to the low scores. The relationship pictured in this chart corresponds to a correlation coefficient of .49.

*The Problem of the Criterion.* We said above that predictive validity can be estimated by determining the correlation between test scores and a suitable criterion measure of success on the job. The joker here is the phrase "suitable criterion measure." One of the most difficult problems that the personnel psychologist or educator faces is that of locating or creating a satisfactory measure of job success to serve as a criterion measure for test validation. It may appear to the student that it should be a simple matter to decide upon some measure of rate of production or some type of rating by superiors. It may also seem that this measure, once decided upon, should be obtainable in an easy and straightforward fashion. Unfortunately, this is not so. Finding or developing acceptable criterion measures usually involves the research worker in the field of tests and measurements in a number of troublesome problems.

Difficulties in obtaining satisfactory criterion measures arise from a variety of sources. There are many types of jobs, such as those of physician, teacher, secretary, or stock clerk, that yield no objective

\* This is not entirely true. What a test "looks like" may be of importance in determining its acceptability and reasonableness to those who will be tested. Thus, a group of would-be pilots may be more ready to accept an arithmetic test dealing with wind drift and gasoline consumption than they would the same essential problems phrased in terms of costs of crops or of recipes for baking cakes. This appearance of reasonableness is sometimes spoken of as "face validity."

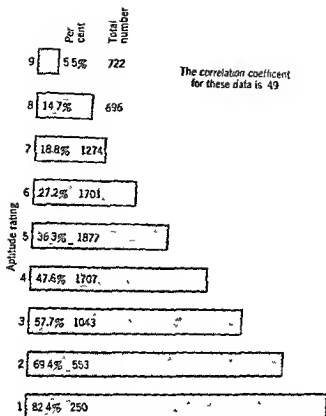


Fig. 7.1. Per cent of cadets eliminated from pilot training at each aptitude level.

record of performance or production. But even when such records are available, they are often influenced by a variety of factors outside the worker's control. Thus, the production record of a weaver may depend not only upon his own skill in threading or adjusting the loom but also on the condition of the equipment, the adequacy of the lighting where he must work, or the color of the thread he must weave. The sales of an insurance agent are not only a function of his own effectiveness as a salesman but also of the territory in which he must work and the supervision and assistance he receives. The problems of effective rating of personnel are discussed in detail in Chapter 13. It suffices to indicate here that ratings are often unstable and influenced by many factors other than the proficiency of the person being rated.

There are always many criterion measures that might be obtained and used for validating a selection test. In addition to quantitative performance records and subjective ratings, which have already been mentioned, we might use later tests of proficiency. This is the type

of situation that is involved when a college entrance mathematics test is validated in terms of its ability to predict later performance on a comprehensive examination on college mathematics. Here the comprehensive examination serves as the criterion measure. Another common type of criterion is grades in some type of educational or training program. Thus, tests for the selection of engineers may be validated against course grades in engineering school.

All criterion measures are only partial in that they measure only a part of success on the job or only the preliminaries to actual job performance. This last is true of the engineering school grades mentioned above. They represent a relatively immediate but quite partial criterion of success as an engineer. The ultimate criterion is some appraisal of the man's lifetime success in his profession. In the very nature of things, such an ultimate criterion is inaccessible to us and we must be satisfied with substitutes for it. These substitutes are only partial and are never completely satisfactory. Our problem is always to choose the most satisfactory from among the measures that it appears feasible to obtain. We are faced, then, with the problem of deciding which of several criterion measures is most satisfactory. How shall we arrive at this decision?

*Qualities Desired in a Criterion Measure.* There are four qualities that we desire in a criterion measure. In order of their importance they are (1) relevance, (2) freedom from bias, (3) reliability, and (4) availability.

We judge a criterion to be relevant to the extent that score on the criterion measure is determined by the same factors that determine success on the job. In appraising the relevance of a criterion, we are thrown back once more upon rational considerations. There is no empirical evidence that will tell us whether a particular criterion measure is or is not relevant. For achievement tests we found it necessary to rely upon the best available professional judgment to determine whether the content of the test accurately represented our objectives. In the same way, with respect to a criterion measure it is also necessary to rely upon professional judgment to provide the appraisal of the degree to which any available partial criterion measure is relevant to the ultimate criterion of job success.

A second factor important in a criterion measure is that of freedom from bias. By this we mean that the measure should provide each person with the same opportunity to make a good score. Examples of biasing factors are such things as variation in wealth from one district to another in our previous example of the insurance salesman, variation in the quality of equipment and conditions of work of a

factory worker, variation in generosity of the bosses rating private secretaries, or variation in the skill of teachers instructing pupils in different classes. We can see that it will be difficult to get meaning from the relationship of test results to a criterion score if that score depends upon factors in the conditions of work rather than factors in the individual worker.

The topic of reliability will be discussed in general terms later in this chapter. As it applies to the criterion scores, the problem is merely this: a measure of success on the job must be stable or reproducible if it is to be predicted by any type of test device. If the criterion performance is one that jumps around from day to day so that the person who shows high job performance one week may show low job performance the next, then there is no possibility of finding a test that will predict it. A measure that is fundamentally unstable itself cannot be predicted by anything else.

Finally, in the choice of criterion measure one always encounters practical problems of convenience and availability. How long is it going to take to get a criterion score for each individual? How much is it going to cost? Though a personnel research program can often afford to spend a substantial part of its effort in getting good criterion data, there is always a practical limit. Any choice of a criterion measure must take this practical limit into account.

#### THE INTERPRETATION OF VALIDITY COEFFICIENTS

Suppose that we have gathered test and criterion scores for a group of individuals and computed the correlation between them. Perhaps our predictor is a scholastic aptitude test, and the criterion is an average of college freshman grades. How shall we now decide whether the test is a good predictor?

Obviously, other things being equal, the higher the correlation, the better. In one sense, our only basis for evaluating any predictor is in relation to other possible prediction procedures. Does test A yield a higher or lower validity coefficient than other tests? Than other types of information, such as high-school grades or rating by school principals? We will look with favor on any measure whose validity for a particular criterion is higher than that of measures previously available to us.

Some representative validity coefficients are exhibited in Table 7.1. These give some picture of the size of correlation that has been obtained in previous work of different kinds. The investigator concerned with a particular course of study or a particular job criterion will, of course, need to become intimately acquainted with validities found for his particular criterion.

Table 7.1. Validity of Selected Tests as Predictors of Certain Educational and Vocational Criteria

Predictor Test	Criterion Variable	Validity Coefficient
<i>Pintner General Ability Test</i>	<i>Metropolitan Achievement—Reading Comp. (Gr. 5)</i>	.76
	<i>Metropolitan Achievement—Total Score (Gr. 5)</i>	.84
<i>ACE Psychological Exam—L Score</i>	College Grades—English	.48
	College Grades—Math	.33
	College Grades—Art	.24
<i>Seashore Tonal Memory Test</i>	Performance test on stringed instrument	.28
<i>Short Employment Test</i>		
Word Knowledge Score	Production index—80 bookkeeping machine operators	.10
Word Knowledge Score	Job grade—106 stenographers	.53
Arithmetic Skill Score	Production index—80 bookkeeping machine operators	.26
Arithmetic Skill Score	Job grade—106 stenographers	.60
<i>Differential Aptitude Tests</i>		
<i>Verbal Reasoning</i>	English grades 3½ years later	.57
<i>Space Relations</i>	English grades 3½ years later	.01
<i>Mechanical Reasoning</i>	English grades 3½ years later	.17

The usefulness of a test as a predictor depends not only on how well it correlates with a criterion, but also on how much *new* information it gives. Thus, the *Differential Aptitude Tests' Verbal Reasoning Test* correlates on the average .48 with high-school English grades, and a test of sentence usage correlates .51 with the same grades. But the two tests have an intercorrelation of .62. They overlap and, in part at least, the information each test provides is the same as that provided by the other test. The net result is that pooling the two tests can give a validity coefficient of no more than .55. If the two tests were uncorrelated, each giving evidence completely independent of the other, the combination of the two tests would give a validity coefficient of .70.\*

\* Statistical procedures have been developed that enable us to determine the best weighing to give the two or more predictors and to calculate the correlation that will result from this combination. The procedures for computing the weights for the separate components (called regression weights) and the correlation (multiple correlation) resulting from them are beyond the scope of this discussion but will be found in standard statistics texts.

Clearly, the higher the correlation between a test or other predictor and a criterion, the more pleased we shall be. But in addition to this relative standard, we should like some absolute one. How high must the validity coefficient be for the test to be useful? What is a "satisfactory" validity? This is a little bit like asking, "How high is up?" However, we can try to give some sort of answer.

To an organization using a test as a basis for deciding whether to hire a particular job applicant or admit a particular student, the significant question is: How much more often will we make the right decision on whom to hire or admit if we use this test than if we operate on a purely chance basis or on the basis of some less valid measure? The answer to this question depends in considerable measure on the proportion of individuals who must be accepted. A selection procedure can do much more for us if we need to accept only the individual who appears to be the best one in every ten applicants than if we must accept nine out of ten. However, to provide a specific example, let us assume that we will accept half of the applicants. We may then ask what per cent of the ones we accept will fall in the upper half of the whole group in job success, i.e., in what per cent of our decisions do we make a "correct" choice? The per cent of correct choices that will result for correlations of different sizes is shown in Table 7.2.

Table 7.2. Per Cent of Correct Assignments When 50 Per Cent of Group Must Be Selected

Validity of Selection Procedure	Per Cent of Correct Choices
	50.0
.00	56.4
.20	63.1
.40	66.7
.50	70.5
.60	74.7
.70	79.5
.80	85.6
.90	

Table 7.2 indicates that when the correlation is zero, the per cent of correct decisions is 50. This is exactly the chance value. Fifty per cent of our cases are defined as successes, i.e., as falling in the upper half of the total group, and if we had picked our students or employees by just flipping a coin, we could have been right 50 per cent of the time. The improvement in our "batting average" as the correlation goes up is shown in the table. Thus, for a correlation



of .40 we will pick right 63.1 per cent of the time; with a correlation of .80 our percentage will be 79.5, and so forth.

The table shows not only our accuracy for any given correlation but our gain in accuracy if we raise the validity of our predictor. Thus, if we were able to replace a predictor with a validity of .40 by one with a validity of .60, we would increase our per cent of correct decisions from 63.1 to 70.5. All these percentages refer, of course, to the ground rules set in the previous paragraph. However, Table 7.2 gives a fairly representative basis for understanding the effects of a selection program from the point of view of the employing or certifying agency.

In many selection situations, the gain can be crudely translated into a dollars-and-cents saving. Thus, if it costs a company \$500 to employ and train a new worker up to the point of useful productivity, a selection procedure that raised the per cent of successes from 56.4 to 63.1 would yield a saving in wasted training expenses alone of \$3350 per 100 men tested. This takes no account of the possibility that the test-selected men might also be *better* workers after they had completed their training.

Another way of appraising the practical significance of a correlation coefficient, and one that is perhaps more meaningful from the point of view of the person being tested, is shown in Table 7.3. The rows in the little tables represent the fourths of a group of applicants, potential students or employees, with respect to a predictor test. The columns indicate the per cent of cases falling in each fourth on the criterion score. Look at the little table in Table 7.3 corresponding to a validity coefficient of .50. We see that of those who fall in the lowest fourth on our predictor 480 out of 1000 or 48.0 per cent fall in the lowest fourth on the criterion score, 27.9 per cent in the next lowest fourth, 16.8 per cent in the next to highest fourth, and 7.3 per cent in the highest fourth. The diagonal entries represent cases that fall in the same fourth on both predictor and criterion. The further we get from the diagonal, the greater the discrepancy between prediction and outcome.

This table emphasizes not so much the gain from using the predictor test as the variation in job success of those who are similar in predictor scores. From the point of view of schools or employers, the important thing is the improved percentage of accuracy illustrated in Table 7.2. Dealing in large numbers, they can count on gaining from any predictor that is more valid than the procedure currently in use. From the point of view of the single individual, the many marked discrepancies between predicted and actual success shown in Table 7.3 may

Table 7.3. Accuracy of Prediction for Different Values of the Correlation Coefficient  
(1000 cases in each row or column)

$r = .00$					$r = .60$				
Quarter on Predictor	Quarter on Criterion				Quarter on Predictor	Quarter on Criterion			
	4th	3rd	2nd	1st		4th	3rd	2nd	1st
1st	250	250	250	250	1st	45	141	277	537
2nd	250	250	250	250	2nd	141	264	318	277
3rd	250	250	250	250	3rd	277	318	264	141
4th	250	250	250	250	4th	537	277	141	45

$r = .40$					$r = .70$				
Quarter on Predictor	Quarter on Criterion				Quarter on Predictor	Quarter on Criterion			
	4th	3rd	2nd	1st		4th	3rd	2nd	1st
1st	104	191	277	428	1st	22	107	270	601
2nd	191	255	277	277	2nd	107	270	353	270
3rd	277	277	255	191	3rd	270	353	270	107
4th	428	277	191	104	4th	601	270	107	22

$r = .50$					$r = .80$				
Quarter on Predictor	Quarter on Criterion				Quarter on Predictor	Quarter on Criterion			
	4th	3rd	2nd	1st		4th	3rd	2nd	1st
1st	73	168	279	480	1st	6	66	253	675
2nd	168	268	295	279	2nd	66	271	410	253
3rd	279	295	268	168	3rd	253	410	271	66
4th	480	279	168	73	4th	675	253	66	6

seem at least as important. If he has done poorly on the test, an applicant may be less impressed by the fact that the probability is that he will be below average on the job than by the fact that he may do very well. He may always be the exception.

One further point can well be emphasized in conclusion. Validity is always specific to a particular curriculum or a particular job. When an author or publisher claims that his test is valid, it is always appropriate to ask: Valid for what? A test in social studies that accurately

represents the content and objectives of one program of instruction may be quite inappropriate for the program in a different community. The test must always be evaluated against the objectives of a specific program of instruction. Again, a test quite valid for picking department store sales clerks who will be pleasant to customers, informed about their stock, and accurate in financial transactions may be entirely useless in identifying effective insurance salesmen who will go out and find or create new business. Validity must always be evaluated in relation to the specific situation in which a measure is to be used.

### VALIDITY AS SIGNIFYING

Sometimes we ask, with respect to a psychological test, neither "How well does this test predict job success?" nor "How well does this test represent our curriculum?", but "What does this test *mean* or *signify*?" What does the score tell us about an individual? Does it correspond to some meaningful trait or construct that will help us in understanding him? For this question of whether the test tells us something meaningful about people the term *construct* validity has been used.

Let us examine one specific testing procedure and see how its validity as a measure of a useful psychological quality or construct was studied. McClelland<sup>6</sup> developed a testing procedure to appraise the individual's need or motivation to achieve—to succeed and do well. The test used pictures like those in the *Thematic Apperception Test* (see Ch. 15). The individual was called upon to make up a story about each picture, telling what was happening and how it turned out. A scoring system was developed for these stories, based on counting the frequency with which themes of accomplishment, mastery, success, and achievement appeared in the story material. Thus, each individual received a score representing the strength of his motivation to achieve. Now, how are we to determine whether this measure has validity in the sense of truthfully describing a meaningful aspect of the individual's make-up? Let us see how McClelland and his co-workers proceeded.

In essence, the investigators proceeded to ask: "With what should a measure of achievement motivation be related?" They made a series of predictions. Some of the predictions were as follows:

1. Those high on achievement motivation should do well in college, in relation to their scholastic aptitude.
2. Achievement motivation should be higher just after students have been taking tests described to them as measuring their intelligence.

3. Those high on achievement motivation should complete more items on a motivated speeded test.

4. Achievement motivation should be higher for children of families emphasizing early independence.

Each of these predictions was based on a sort of "theory of human behavior." Thus, academic achievement is seen as a combination of ability and effort. Presumably those with higher motivation to achieve will exert more effort and will, ability being equal, achieve higher grades. A similar chain of reasoning lies back of each of the other predictions.

In general, McClelland found that most of his predictions were supported by the experimental results. The fact that the test scores were related to a number of other events in the way that was predicted from a rational analysis of the trait that the test was presumably measuring lent support to the validity of the test procedure as measuring a meaningful trait or construct, whose essential characteristics are well summarized by the label "achievement motivation."

A great many of our psychological tests, and, to a lesser extent, some educational tests, are intended to measure general traits or qualities of the individual. Verbal reasoning, spatial visualizing, sociability, introversion, mechanical interest are all designations of traits or constructs. Tests of these functions are valid insofar as they behave in the way that such a trait should reasonably be expected to behave. Some of the indicators of how a trait (and therefore a test of it) should behave are:

1. Its correlations with other tests, especially tests that are already accepted measures of the function in question. Thus, many group intelligence tests have been validated in part by their correlations with earlier tests, and especially with the individually administered *Stanford-Binet*.

2. Its correlations with outside facts about the individual, and its ability to differentiate between different groups. Thus, the fact that score on achievement need on the *Edwards Personal Preference Schedule* is higher for those with more education, those with higher incomes, those from urban rather than rural backgrounds, and those in their 30's rather than older groups seems consistent with the predictions that we would make, and supports the validity of the score.

3. Its response to changes in external conditions, especially to conditions that are experimentally induced for the specific purpose of testing the responsiveness of the instrument. Thus, flicker fusion has been proposed as an indicator of anxiety. One study<sup>1</sup> compared flicker

## ✓ RETEST WITH THE SAME TEST

If we wish to find how reliably we can evaluate an individual's weight, we can have him weighed twice. It may be a reasonable precaution to have the two measures taken independently by two persons. We don't want the experimenter's recollection of the first score to color the second score. It may be desirable to have the two weighings done on different days. That depends on what we are interested in. If we want to know how accurately we can carry out the process of weighing a person, the two measures should be carried out one right after the other. Then we know that the *person* has stayed the same and that the only source of variation or "error" is in the operation of weighing him. If we want to know how precisely a given weight characterizes a person from day to day—how closely we can predict his weight next week from what he weighs today, it would be appropriate to measure him on two separate occasions. Now we are interested in *variation within the individual* as well as *variation due to the operation of measurement*.

Sometimes we are interested in variation within the individual; sometimes we are not. We may ask: How accurately does our measurement characterize S at this moment of time? Or we may ask: How accurately does our measure of S today describe him as he will be tomorrow, or next week, or next month? Both are sensible questions. But they are not the same question. The data we must gather to answer one are different from the data we need to answer the other.

To study the reliability of such a physical characteristic of a person as weight, repetition of the measurement is a straightforward and satisfactory operation. It appears satisfactory and applicable also with some simple types of behavior, such as speed of reaction or muscular strength. But suppose now we are interested in the reliability of a test of reading comprehension. Let us assume that the test is made up of six reading passages with ten questions on each. We administer the test once and then immediately administer it again. What happens? Certainly, the child is not going to have to reread all the material he has just read. He may do so in part, but to a considerable extent his answers the second time will involve merely remembering what answer he had chosen the time before and marking it again. If he had not been able to finish the first time, he will now be able to work ahead and spend most of his time on new material. These same effects will hold true to some degree even over a longer period of time. Clearly, this sort of test given a second time does not present the same task that it did the first time.

There is a second consideration entering into the repetition of such

a test as a reading comprehension test. Suppose that one of the five passages in the test was about baseball and that a particular boy was an expert on baseball. The passage would then be especially easy for him, and he would in effect get a bonus of several points. The test would overestimate his general level of reading ability. But note that it would do it consistently on both testings because the material remains the same. The error for individual S is a *constant error* in the two testings. Since it affects both his scores in the same way, it makes the test look reliable rather than unreliable.

In such an area of ability as reading, we must recognize the possibility that an individual does not perform uniformly well throughout the whole area. His specific interests, experiences, and background give him strengths and weaknesses. A particular test is *one sample* from the whole area. How well individual S does on the test, relative to others, is likely to depend in some degree upon the particular sample of tasks chosen to represent the area of ability or personality we are trying to appraise. If the sample remains the same for both measurements, his behavior will stay more nearly the same than if the sample of tasks is varied.

Note that so far we have identified three main sources of variation in performance that will tend to reduce the precision of a particular score as a description of an individual:

1. Variation in response to the test at a particular moment in time.
2. Variation in the individual from time to time.
3. Variation arising out of the particular sample of tasks chosen to represent an area of behavior.

Retesting the individual with identically the same test can be arranged to reflect the first two types of "error," but this procedure cannot evaluate the effects of the third type. In addition, there may be the memory and practice effects to which we referred above.

#### ✓ PARALLEL TEST FORMS

Concern about this third source of variation, variation arising because of the necessity of choosing a particular sample of tasks to represent a whole area of behavior, leads us to another set of procedures for evaluating reliability. If the sampling of items may be a significant source of "error," and if, as is usually the case, we want to know with what accuracy we may generalize from the specific score to the area of behavior it is supposed to represent, we must develop some procedures that take account of this variation due to the sample of tasks. We may do this by correlating two equivalent forms of a test.

Equivalent forms of a test should be thought of as forms built according to the same specifications but composed of separate samples of behavior in the defined area. Thus, two equivalent reading tests should contain reading passages and questions of the same difficulty. The same sorts of questions should be asked, i.e., the same balance of specific fact and general idea questions. The same types of passages should be represented, i.e., expository, argumentative, esthetic. But the specific passages and questions should be different.

If we have two forms of a test, we may give each pupil first one form and then the other. They may follow each other immediately if we are not interested in stability over time, or may be separated by an interval if we are. The correlation between the two forms will provide an appropriate reliability coefficient. If a time interval has been allowed between the testings, all three of our sources of variation will have had a chance to get in their effects—variation arising from the measurement itself, variation in the individual over time, and variation due to the sample of tasks.

To ask that a test yield consistent results under these conditions is the most rigorous standard we can set for it. And if we want to use our test results to generalize about what Johnny will do on other tasks of this general sort next week and next month, then this is the appropriate standard by which to evaluate a test. For most educational situations, this is the way we want to use test results, and so evidence based on equivalent test forms should usually be given the most weight in evaluating the reliability of a test.

The use of two parallel test forms provides a very sound basis for estimating the precision of a psychological or educational test. This procedure does, however, raise some practical problems. It demands that two parallel forms of a test be available and that time be allowed for administering two separate tests. Sometimes no second form of a test exists, or no time can be found for a second testing. To administer a second separate test is often likely to represent a somewhat burdensome demand upon available resources. These practical considerations of convenience and expediency have made test makers receptive to procedures that extract an estimate of reliability from administration of only one form of a test. However, such procedures are compromises at best. The correlation between two parallel forms, usually administered with a lapse of several days or weeks in between, represents the preferred procedure for estimating reliability.

#### — SUBDIVIDED TEST

The most widely used procedure for estimating reliability from a single testing divides a particular test up into two presumably equiv-

alent halves. The half-tests may be assembled on the basis of careful examination of the content and difficulty of each item, making a systematic effort to balance out the content and difficulty level of the two halves. A simpler procedure, which is often relied upon to give equivalent halves, is to put alternate items into the two half-tests, that is, to put all the odd-numbered items in one half-test and all the even-numbered items in the other. This is usually a sensible procedure, since items of similar form, content, or difficulty are likely to be grouped together in a test. For a reasonably long test, say, of 60 items or more, splitting the test up in this way will tend to balance out factors of item form, content covered, and difficulty level. The two half-tests will have a good probability of constituting "equivalent" tests, as these are defined in the preceding section.

The procedures we are discussing now divide the test in half only for scoring, not for administration. That is, a single test is given at a single sitting and with a single time limit. However, two separate scores are derived—one by scoring the odd-numbered items and one by scoring the even-numbered items. The correlation between these two scores provides a measure of the accuracy with which the test is measuring the individual.

However, it must be noted that the computed correlation is between two half-length tests. This value is not directly applicable to the full-length test, which is the actual instrument prepared for use. In general, the larger the sample of a person's behavior we have, the more reliable the measure will be. The more behavior we record, the less our measure will depend upon chance elements in behavior of the individual or in the particular sampling of tasks. Single lucky answers or momentary lapses of attention will be more nearly evened out.

Where the two halves of the test, which gave the scores actually correlated, are equivalent, we can get an unbiased estimate of total-test reliability from the correlation between the two half-tests. This estimate is given by the formula

$$r_{11} = \frac{2r_{\frac{1}{2}\frac{1}{2}}}{1 + r_{\frac{1}{2}\frac{1}{2}}} \quad (1)$$

where  $r_{11}$  is the estimated reliability of the full-length test,  
 $r_{\frac{1}{2}\frac{1}{2}}$  is the actual correlation between two half-length tests.

Thus, if the correlation between the two halves of a test is .60, formula 1 would give

$$r_{11} = \frac{2(0.60)}{1 + 0.60} = \frac{1.20}{1.60} = .75$$



This formula, referred to generally as the Spearman-Brown Prophecy Formula from the names of its originators and function, makes it possible for us to compute an estimate of reliability from a single administration of a single test.

The appealing convenience of the split-half procedure has led to its wide use. Many test manuals will be found to report this type of reliability coefficient and no other. Unfortunately, this coefficient has several types of limitations, which we must now examine.

In the first place, when we have extracted two scores from a single testing, both scores necessarily represent the individual as he is at the same moment of time. Even events lasting only a few minutes will affect both scores about equally. In other words, variation of the individual from day to day cannot be reflected in this type of reliability coefficient. It can only give evidence as to the precision with which we can appraise him at a specific moment in time.

In the second place, a split-half reliability coefficient becomes meaningless when a test is highly speeded. Suppose we have a test of simple arithmetic, made up of problems like  $3 + 5 = ?$ , and that the test is being used with adults with a 2-minute time limit. We will get wide differences in score on such a test, but the differences will be primarily differences in speed. Errors will be a minor factor. The person who gets a score of 50 will very probably have attempted just 50 items, *and of these 25 will be odd and 25 will be even*. In other words, the two halves of the test will appear perfectly consistent, because opportunity to attempt items is automatically balanced out for the two half-tests.

Few tests depend as completely upon speed as does the one that we have chosen to illustrate our point. However, many involve some degree of speeding. This speed factor will tend to inflate estimates of reliability based on the split-half procedure. The amount of over-estimation will depend upon the degree to which the test is speeded, being greater for those tests in which speed plays a greater role. However, speed enters in sufficiently generally so that split-half estimates of reliability should always be discounted. Test users should demand that commercial publishers provide reliability estimates based on parallel forms of the test.

#### RELIABILITY ESTIMATED FROM ITEM STATISTICS

The teacher or investigator who makes much use of tests and who reads extensively in test manuals will encounter one other type of procedure for estimating test reliability from a single test administra-

tion. This procedure, also named for its originators, yields what is referred to as a Kuder-Richardson reliability coefficient. The essential assumption in the procedure is that the items within one form of a test have as much in common with one another as do the items in that one form with the corresponding items in a parallel or equivalent form. This means that the items in a test are homogeneous in the sense that every item measures the same general factors of ability or personality as do the others. If this assumption is sound, the Kuder-Richardson procedure leads to a reliability estimate that has essentially the same interpretation as the odd-even coefficient we have just considered. The Kuder-Richardson estimate likewise (1) takes no account of variation in the individual from time to time, and (2) is inappropriate for speeded tests. Within these two limitations, it provides a conservative estimate of the split-half type of reliability.\*

#### COMPARISON OF METHODS

A summary comparison of the different procedures for estimating reliability is given in Table 7.4. This shows four factors that may make a single test score an inaccurate picture of the individual's usual performance. The table shows which of the factors are represented in each of the procedures for estimating reliability we have discussed. It can be seen that the different procedures are not equivalent. Only administration of parallel test forms with a time interval between permits all sources of variation to have their effects. Each of the other

\* A widely used form of the Kuder-Richardson procedure (their Formula 20) takes the form

$$r_{11} = \left( \frac{n}{n-1} \right) \left( \frac{s_t^2 - \sum pq}{s_t^2} \right)$$

where  $r_{11}$  is the estimate of reliability.

$n$  is the number of items in the test.

$s_t^2$  is the standard deviation of the test.

$\sum$  means "take the sum of" and covers the  $n$  items.

$p$  is the per cent passing a particular item.

$q$  is the per cent failing the same item.

A formula involving simpler calculations (their Formula 21), which yields a reasonably close approximation to the above, is

$$r_{11} = \frac{n}{n-1} \left[ 1 - \frac{M_t \left( 1 - \frac{M_t}{n} \right)}{s_t^2} \right]$$

where  $M_t$  is the mean score of the group and the other symbols have the same meaning as given above.

Table 7.4. Sources of Variation Represented in Different Procedures for Estimating Reliability

Sources of Variation	Experimental Procedure for Estimating Reliability					
	Immediate Retest, Same Test	Retest after Interval, Same Test	Parallel Test Form without Time Interval	Parallel Test Form with Time Interval	Odd-Even Halves of Single Test	Kuder-Richardson Analysis, Single Test
<i>How much the score can be expected to fluctuate owing to:</i>						
Variations arising within the measurement procedure itself	X	X	X	X	X	X
Changes in the individual from day to day		X		X		
Changes in the specific sample of tasks			X	X	X	X
Changes in the individual's speed of work	X	X	X	X		

methods masks some source of variation that may be significant in the actual use of tests. Retesting with the same identical test neglects variation arising out of the sample of items. Whenever all the testing is done at one point in time, variation of the individual from day to day is neglected. When the testing is done as a unit with a single time limit, variation in speed of responding is neglected. The facts brought out in this table should be borne in mind in evaluating reliability data found in a test manual or in the report of a research study.

#### INTERPRETATION OF RELIABILITY DATA

Analysis of data obtained from a general intelligence test for elementary-school children has yielded a reliability coefficient of .85. How shall we interpret this result? What does it mean concerning the precision of an individual's score? Should we be pleased or dissatisfied to get a coefficient of this size?

We have already tried to give some content and meaning to correlation coefficients in Fig. 5.7 and in Tables 5.8, 7.1, 7.2, and 7.3. These have shown typical values of the correlation coefficient, the scatter of scores for representative correlations, and the accuracy of prediction with correlations of different sizes. A further contribution to the interpretation of test reliability is found in the relationship between the reliability coefficient and the standard error of measurement.

It will be remembered that the standard error of measurement is an estimate of the standard deviation that would be obtained for a series of measurements of the same individual. (It is assumed that

he is not changed by being measured.) The standard error of measurement can be calculated from the reliability coefficient by the formula

$$s_m = s_t \sqrt{1 - r_{11}} \quad (2)$$

where  $s_m$  is the standard error of measurement.  
 $s_t$  is the standard deviation of test scores.  
 $r_{11}$  is the reliability coefficient.

Suppose that our test has a reliability of .85 and a standard deviation of 15 points. Then we have

$$s_m = 15\sqrt{1 - .85} = 15\sqrt{.15} = 5.7$$

In this instance, a set of measures of a particular person would have a standard deviation of 5.7 points. Remember that a fairly uniform proportion of observations fall within any given number of standard deviation units from the mean. Certain values for this relationship were given in Table 5.6. This table shows that for a normal curve 31.8 per cent of cases, or about 1 in 3, differ from the mean by as much as 1 standard deviation; 4.6 per cent by as much as 2 standard deviations. Applying this to our case, in which the standard deviation of our measurements is 5.7 points, we could say that there is about 1 chance in 3 that a score that we get for an individual differs from his "true" score by as much as 5.7 points (1 standard error of measurement). There is about 1 chance in 20 that it differs by as much as 11.4 points (2 standard errors of measurement).

The values shown above are fairly representative of what might be found for intelligence quotients from one of the commercially distributed group intelligence tests applied to children in the upper elementary grades. Note that even with this relatively high reliability coefficient, appreciable errors of measurement are possible in at least a minority of cases. Shifts of 5 or 10 points of IQ can be expected fairly frequently just because of errors of measurement. Anyone who is impressed by and tries to interpret an IQ difference of 5 points between two persons or two testings of the same person has been fooled into thinking the test has a precision that it simply does not possess. Further testing could perfectly well reverse the result. Any test score or comparison of test scores must be made with acute awareness of the standard error of measurement.

The manner in which the standard error of measurement is related to the reliability coefficient is shown in Table 7.5. We see that the magnitude of errors decreases as the reliability increases, but we also

Table 7.5. Standard Error of Measurement for Different Values of Reliability Coefficient

Reliability Coefficient	Standard Error of Measurement	
	General Expression	When $S_t^* = 10$
.50	$.71 S_t^*$	7.1
.60	$.63 S_t^*$	6.3
.70	$.55 S_t^*$	5.5
.80	$.45 S_t^*$	4.5
.85	$.38 S_t^*$	3.8
.90	$.32 S_t^*$	3.2
.95	$.22 S_t^*$	2.2
.98	$.14 S_t^*$	1.4

\*  $S_t$  signifies the standard deviation of the test.

see that errors of appreciable size will still be found even with reliability coefficients of .90 or .95. In interpreting the score of a particular individual, it is the standard error of measurement that must be kept in mind. If we think of a range extending from 2 standard errors of measurement above the obtained score to 2 below, we will have a band within which we can be reasonably sure (19 chances in 20) that the individual's true score lies. Thus, in the case of the intelligence test described in previous paragraphs, we can think of a test IQ of 90 as meaning rather surely an IQ lying between about 80 and 100. If we think in those terms, we shall be much more discreet in interpreting and using test results.

When interpreting the test score of an individual, it is desirable to think in terms of the standard error of measurement and to be somewhat bumble and tentative in drawing conclusions from that test score. But for making comparisons between tests and for a number of types of test analysis, the reliability coefficient will be more useful. Where measures are expressed in different units, as height in inches and weight in pounds, the reliability coefficient provides the only possible basis for comparison. Since the competing tests in a given field, such as primary reading, are likely to use types of scores that are not really comparable, the reliability coefficient will usually represent the only satisfactory basis for test comparison. *Other things being equal*, we shall prefer the test with the higher reliability coefficient, that is, the test that provides a more consistent ranking of the individual within his group.

The other things that may not be equal are primarily considerations of validity and practicality. Validity, in so far as we can appraise it, is the crucial test of a measurement procedure. Reliability is important only as a necessary condition for a measure to have validity. The ceiling for the possible validity of a test is set by its reliability. A test must measure *something* before it can measure what we want it to measure. A measuring device with a reliability of .00 is reflecting nothing but chance factors. It does not correlate with itself and cannot correlate with anything else. The theoretical ceiling for the validity coefficient of a test (i.e., its correlation with some criterion measure representing success in learning or on the job) is the square root of its reliability coefficient. Thus, a test with reliability coefficient of .36 could not give a validity coefficient above .60, and one with a reliability coefficient of .64 could not possibly yield a validity coefficient above .80. Only to the extent that a test measures something accurately can it measure it validly.

The converse of the relationship we have just presented does not follow. A test may measure with the greatest precision and still have no validity for our purposes. Thus, we can measure head size with a good deal of accuracy, but the measure is still useless as an indicator of intelligence. Validity is something over and beyond mere accuracy of measurement.

Considerations of cost, convenience, etc. may also sometimes lead to a decision to use a less reliable test. We may accept a less reliable 40-minute test in preference to a more reliable 3-hour one because the 3 hours of testing time is too much of a burden in view of the purpose the test is designed to serve.

Within the limitations discussed in the preceding paragraphs, we shall prefer the more reliable test. There are several factors that must be taken into account, however, before we can fairly compare the reliability coefficients of two or more different tests. These will be discussed in the paragraphs that follow.

1. *Range of the Group.* The reliability coefficient indicates how consistently a test places each individual relative to the others in the group. When there is little shifting from test to retest or from A to form B, the reliability coefficient is high and vice versa. But the extent to which individuals will switch places depends on how closely similar they are. It does not take very accurate testing to differentiate the reading ability of second graders from that of seventh graders. But to place each second grader accurately within his own class is much more demanding.

If children from several different grades are pooled together, we may expect a much higher reliability coefficient. For example, the manual for the *Otis Quick-Scoring Mental Ability Test*—Beta reports alternate-forms reliabilities for single grade groups ranging from .65 to .87. The average value is .78. But pooling the complete range of grades (4–9), the reliability coefficient is reported as .96. These data are all for the same test. They reflect the same precision. Yet the coefficient for the combined groups is strikingly higher. Similar data are reported for the *Durrell-Sullivan Reading Achievement Test*. The data in this case involve a range of four grades—from grade three through grade six. Reliability coefficients are split-half reliabilities based on a single testing. In the case of the *Word Meaning Test*, the average coefficient for a single grade is .93, whereas the correlation for all four grades together is .97. For the test of *Paragraph Meaning* the corresponding values are .87 and .94.

In evaluating a reported reliability coefficient, the range of ability in the group tested must be taken into account. If the reliability coefficient is based upon a combination of age or grade groups, it must usually be sharply discounted, as can be seen above. But even in less extreme cases, account must be taken of the variability of talent within the group. Reliabilities for age groups will tend to be somewhat higher than for grade groups, because an age group will usually contain a greater spread of talent than a single grade. A sample made up of children from a wide range of socio-economic levels will tend to yield higher reliabilities than a very homogeneous one. In comparing different tests, one must take account of the type of sample on which the reliability data were based, in so far as this can be determined from the reported facts, and judge more severely the test whose reliability is based on the more heterogeneous group.

2. *Level of Ability in the Group.* Precision of measurement by a test may be related to the ability level of the persons being measured. However, no simple rule can be formulated for stating the nature of this relationship. It depends upon the way in which the particular test was built. For those people for whom the test is very hard, so that they are doing a large amount of guessing, accuracy is likely to be low. At the other extreme, if a test is very easy for a group, so that all of them can do most of the items very easily, it may be expected to be ineffective in discriminating among the members of the group. When everyone can do the easy items, it is as if we had shortened the test to just the few harder items that some can do and some cannot.

It is possible, also, that a test may vary in accuracy at different

intermediate difficulty levels. The meticulous test constructor will report the standard error of measurement for his test at different score levels. When separate values of the standard error of measurement are reported in the manual, they provide a basis for evaluating the precision of the test for different types of groups. They permit a more appropriate estimate of the accuracy of a particular individual's score. Each individual's score can be interpreted in relation to the standard error of measurement for scores of that level. For example, from data provided by Terman and Merrill \* the standard error of measurement for the 1937 edition of the *Stanford-Binet* for different IQ levels is found to be as follows:

<i>IQ Level</i>	<i>Standard Error of IQ</i>
130 and over	5.2
110-129	4.9
90-109	4.5
70-89	3.8
Below 70	2.2

For this test, the variation that may be expected from one testing to another is very much higher for children with average and above average IQ's than for the retarded child. In the case of the *Wechsler Intelligence Scale for Children*, the standard error of measurement depends upon the age of the group tested. The manual reports values as follows:

7½-year-olds	4.2 points of IQ
10½ " "	3.4 " " "
13½ " "	3.7 " " "

The test is most accurate for an age group in the middle of the age range for which it was intended.

3. *Length of Test.* As we saw on p. 179 in discussing the split-half reliability coefficient, test reliability depends on the length of the test. If we can assume that the quality of the test items and the nature of the examinees remain the same, then the relationship of reliability to length can be expressed by a simple formula. The formula is

$$r_{nn} = \frac{nr_{11}}{1 + (n-1)r_{11}} \quad (3)$$

where  $r_{nn}$  is the reliability of a test  $n$  times as long as the original test.

$r_{11}$  is the reliability of the original test.

$n$  is, as indicated, the factor by which the length of the test is increased.

This is a more general form of formula 1 found on p. 179.



Suppose we have a spelling test made up of 20 items which has a reliability of .50. We want to know how reliable the test will be if it is lengthened to contain 100 items comparable to the original 20. The answer is

$$r_{nn} = \frac{5(.50)}{1 + 4(.50)} = \frac{2.50}{3.00} = .83$$

As the length of the test is increased, the chance errors of measurement more or less cancel out; score comes to depend more and more completely upon the characteristics of the person being measured; and a more accurate appraisal of him is obtained.

Of course, how much we can lengthen a test is limited by a number of practical considerations. It is limited by the amount of time available for testing. It is limited by factors of fatigue and boredom on the part of examinees. It is sometimes limited by the stock of good test items that it is possible to construct. But within these limits, reliability can be increased as needed by lengthening the test.

One special type of lengthening is represented by increasing the number of raters who rate an individual or a product he has produced. If several raters of equal competence or equal familiarity with the ratee are available, a pooling of their ratings will produce increased reliability in the composite rating, and this increase will be described by the same formula we have just been considering.

*4. Operations Used for Estimating.* How high a value will be obtained for the reliability coefficient depends also upon which of the several possible sets of experimental operations is used to estimate the reliability. We saw in Table 7.4 that the different procedures treat different sources of variation in different ways, and that it is only the use of parallel forms of a test with a period intervening that includes all four sources of variation in "error." That is, this procedure of estimating reliability represents a more exacting definition of the test's ability to reproduce the same score. The individual must then show consistency both from one sample of tasks to another and from one day to another. We have gathered together a few examples that show reliability coefficients for the same test when these were computed by two different procedures. These are shown in Table 7.6.

The two procedures compared in Table 7.6 are correlation of alternate forms and correlation of half-tests made up from a single form. It will be noted that the alternate-forms correlation is lower in every case. This is consistent with our earlier discussion, in which we pointed out that the alternate-forms procedure constitutes a more demanding test of an instrument's precision. The difference between

Table 7.6. Comparison of Reliability Coefficients Obtained from Equivalent Forms and from Fractions of a Single Test

Test	Alternate Forms	Single Test
<i>Otis Quick-Scoring Intelligence Test—Beta</i>	.84	.90
<i>Pintner-Durost Intelligence Test</i>		
Scale 1, Picture Content	.78	.92
Scale 2, Reading Content	.92	.97
<i>Essential High School Content Battery</i>		
Mathematics	.88	.92
Science	.75	.85
Social Studies	.85	.89
English	.86	.90

the two procedures varies from test to test, being as small as .04 in one instance and as large as .14 in another. But in every instance, it is necessary to discount the odd-even correlation.

#### HOW HIGH MUST THE RELIABILITY OF A MEASUREMENT BE?

Obviously, other things being equal, the more reliable our measuring procedure is, the better satisfied we are with it. A question that is often raised is: What is the *minimum* reliability that is acceptable? Actually, there is no general answer to this question. If we *must* make some decision or take some course of action with respect to an individual, we will do so in terms of the best information we have, however unreliable it may be, provided only that the reliability is better than zero. (Of course, here as always the crucial consideration is the validity of the measure.) The appraisal of any new procedure must always be in terms of other procedures with which it is in competition. Thus, a high-school mathematics test with a reliability coefficient of .80 would look relatively unattractive if tests with reliabilities of .85 to .90 were already available. On the other hand, a procedure for judging "leadership" that had a reliability of no more than .60 might look very attractive if the alternative were a set of uncontrolled ratings having a reliability of .45 to .50.

Although we cannot set an absolute minimum for the reliability of a measurement procedure, we can indicate the level of reliability that is required to enable us to achieve specified levels of accuracy in describing an individual or a group. Suppose that we have given a test to two individuals, and that individual A fell at the 75th percentile of the group while individual B fell at the 50th percentile. What is the probability that A would still surpass B if they were tested again?

Table 7.7. Per Cent of Times Direction of Difference Will Be Reversed in Subsequent Testing for Scores Falling at 75th and 50th Percentile

Reliability Coefficient	Per Cent of Reversals with Repeated Test		
	Scores of Single Individuals	Means of Groups of 25	Means of Groups of 100
.00	50.0	50.0	50.0
.40	40.3	10.9	0.7
.50	36.8	4.6	0.04
.60	32.5	1.2	
.70	27.1	0.1	
.80	19.7		
.90	8.7		
.95	2.2		
.98	0.05		

In Table 7.7 the probability is shown for different values of the reliability coefficient. Thus, where the correlation is .00, there is exactly a fifty-fifty chance that the order of our two individuals will be reversed. When the correlation is .50, the probability of a reversal is 1 in 3. For a correlation of .90, there is still 1 chance in 12 that we will get a reversal on repetition of the testing. To have 4 chances in 5 that our difference will stay in the same direction, we require a reliability of about .80.

Table 7.7 also shows the situation when we are comparing two groups of 25. That is, in class A the average fell at the 75th percentile of some larger reference group, whereas in class B the average fell at the 50th percentile. We ask what the probability is that we would get a reversal if the testing were repeated. Here we still have a fifty-fifty chance when the correlation is .00. However, the security of our conclusion increases much more rapidly as the reliability of our test is increased. When the reliability is .50, the probability of reversal is already down to 1 in 20; with a correlation of .70 it is only 1 in 1000. Thus, a test with relatively low reliability will permit us to make useful studies of and draw accurate conclusions about groups, but relatively high reliability is required if we are to have precise information about individuals.

#### RELIABILITY OF DIFFERENCE SCORES

Sometimes we are less interested in single scores than we are in the relationship between scores taken in pairs. Thus, we may be con-

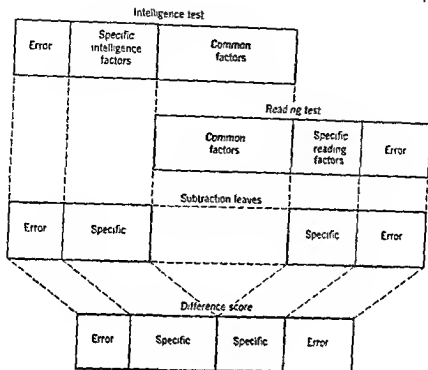


Fig 7.2. Nature of a difference score.

cerned with the differences between scholastic aptitude and reading achievement in a group of pupils, or we may wish to study gains in reading from an initial test given in October to a later test given the following May. In these illustrations, the significant fact for each individual is the difference between two scores. We must inquire how reliable our estimates of these differences are, knowing the characteristics of the two component tests.

It is, unfortunately, true that the appraisal of the difference between two tests usually has substantially lower reliability than the reliability of the two tests taken separately. This is due to two factors: (1) the errors of measurement in both separate tests affect the difference score, and (2) whatever is common to both measures is canceled out in the difference score. We can illustrate the situation by a diagram. Look at Fig. 7.2.

Each bar in Fig. 7.2 represents performance \* on a test, broken up into a number of parts to represent the factors producing this per-

\* More precisely, variance in performance.

formance. The first bar represents an intelligence test, and the second a reading test. Notice that we have divided reading performance into three parts. One part, labeled "common factors," is a complex of general intellectual abilities that operate both in the reading and the scholastic aptitude test. A second part, labeled "specific reading factors," is abilities that appear only in the reading test. The third part, labeled "error," is chance error of measurement. Three similar parts are indicated for the intelligence test. Now look at the third bar, which represents the difference score. In this bar, the common factor has disappeared. It canceled out in our process of subtraction. Only the specific factors and the errors of measurement remain. These are the factors that determine the difference score. And the errors of measurement bulk relatively much larger in this third bar. In the limit, where two tests measured exactly the same common factors, only the errors of measurement would remain in the difference scores, and the differences would have exactly zero reliability.

The reliability of the difference between two scores can be expressed in a fairly simple formula, which reads

$$r_{Diff.} = \frac{\frac{r_{11} + r_{22}}{2} - r_{12}}{1 - r_{12}}$$

where  $r_{11}$  is the reliability of one measure.

$r_{22}$  is the reliability of the other measure.

$r_{12}$  is the correlation between the two measures.

Thus, if the reliability of test A is .80, the reliability of test B is .90, and the correlation of A and B is .60, for the reliability of the difference score we have

$$\begin{aligned} r_{Diff.} &= \frac{\frac{.80 + .90}{2} - .60}{1 - .60} \\ &= \frac{.25}{.40} = .62. \end{aligned}$$

In Table 7.8 the value of  $r_{Diff.}$  is shown for various combinations of values of  $\frac{r_{11} + r_{22}}{2}$  and  $r_{12}$ . Thus, if the average of the reliabilities of our two tests  $\left(\frac{r_{11} + r_{22}}{2}\right)$  is .80, the reliability of the difference score is .80 when the two tests have zero intercorrelation, is .60 when

Table 7.8. Reliability of a Difference Score

Correlation between Two Tests ( $r_{12}$ )	Average of Reliability of Two Tests $\left(\frac{r_{11} + r_{22}}{2}\right)$					
	.50	.60	.70	.80	.90	.95
.00	.50	.60	.70	.80	.90	.95
.40	.17	.33	.50	.67	.83	.92
.50	.00	.20	.40	.60	.80	.90
.60		.00	.25	.50	.75	.88
.70			.00	.33	.67	.83
.80				.00	.50	.75
.90					.00	.50
.95						.00

the intercorrelation is .50, and is .00 when the intercorrelation is .80. It is clear that, as soon as the correlation between the two tests begins to approach the average of their separate reliability coefficients, the reliability of the difference score drops very rapidly.

The low reliability that tends to characterize difference scores is something to which the psychologist and educator must always be sensitive. It becomes a problem whenever he wishes to use test patterns for diagnosis. Thus the judgment that Herbert's reading lags behind his scholastic aptitude is a judgment that must be made a good deal more tentatively than a judgment about either his IQ or his reading grade taken separately. The conclusion that Mary has improved in reading more than Jane must usually be a more tentative judgment than that Mary is now a better reader than Jane. Any difference needs to be interpreted in the light of the standard error of measurement of that difference.\*

Many differences will be found to be quite small relative to their standard error, and are consequently quite undependable. The interpretation of profiles and of gain scores are places where this caution especially applies.

#### EFFECTS OF UNRELIABILITY ON CORRELATION BETWEEN VARIABLES

There is one further effect of unreliability which merits brief attention here because it affects our interpretation of the correlations be-

\*The standard error of measurement of a difference is roughly equal to  $\sqrt{S_{m_1}^2 + S_{m_2}^2}$ , where  $S_{m_1}$  is the standard error of measurement of one test and  $S_{m_2}$  is the standard error of measurement of the other.

tween different measures. Let us think of a measure of reading comprehension and one of arithmetic reasoning. In each of these tests, the individual differences in score are due in part to "true" ability and in part to chance "errors of measurement." But if the errors of measurement are really chance matters, the reading test errors and the arithmetic test errors must be uncorrelated. There is no relationship between one toss of a coin and a later toss of a coin. So we have these uncorrelated errors in the total score. This means that they must water down any correlation that exists between the true scores. That is, the actual scores are a combination of true score and error, so the correlation between actual scores is a compromise between the correlation of the underlying true scores and the .00 correlation that characterizes the errors.

We would like to extract an estimate of the correlation between the underlying true scores from our obtained data in order to understand better how much the functions involved have in common. Fortunately, we can do this quite simply. Such an estimate is provided by the formula

$$r_{1,2_s} = \frac{r_{12}}{\sqrt{r_{11}r_{22}}} \quad (4)$$

where  $r_{1,2_s}$  is the correlation of the underlying "true" scores.

$r_{12}$  is the correlation of the obtained scores.

$r_{11}$  and  $r_{22}$  are the reliabilities of the two measures in question.

Thus, if the correlation between our reading test and arithmetic test is .56, and the reliability coefficients of the tests are respectively .71 and .90, we have

$$r_{1,2_s} = \frac{.56}{\sqrt{(.71)(.90)}} = .70$$

Our estimate is that the correlation between error-free measures of arithmetic and reading would be .70. In thinking of these two functions, it would be appropriate to think of the correlation as .70 rather than .56, though the tests correlate only .56.

#### FACTORS MAKING FOR PRACTICALITY IN ROUTINE USE

Though validity and reliability may be all-important in measures that are to be used for special research purposes, when a test is to be used in classrooms throughout a school or school system a number of down-to-earth practical considerations must also be taken into account.

It is easy for the administrator to pay too much attention to small financial savings or to economies of time that make it possible to fit a test into the standard class period with no shifting of schedules, but, nevertheless, these factors of economy and convenience are real considerations. Furthermore, there are other factors relating to the readiness with which the tests may be given, scored, and interpreted that bear more importantly on the use that will be made of the tests and the soundness of the conclusions that will be drawn from them.

### ECONOMY

The practical significance of dollar savings does not need to be emphasized. Dollars are of very real significance for any educational or industrial enterprise. Economy in the case of tests depends in part on cost per copy. It depends in part on the possibility of using the test booklets over again. From the junior high school on, and possibly even in the upper elementary grades, it is feasible to administer a test using a separate answer sheet. Such a separate answer sheet permits reuse of the test booklets. If a test will be used in successive years or if testing can be scheduled so that different classes or schools will be tested on successive days, an important economy can be effected by using the same test booklet over again several times.

A second aspect of economy is saving of time in test administration. However, this is often false economy. We saw in the previous section that the reliability of a test depends on the length of the test. As far as testing time is concerned, we get about what we give. Some tests may be a little more efficiently designed, so that they give a little more reliable measure per minute of testing time, but, by and large, any reduction in testing time will be accomplished at the price of loss in the precision or the breadth of our appraisal.

A third, and quite significant, aspect of economy is ease of scoring. The clerical work of scoring a battery of tests can become either burdensome if it is done by the already busy teacher or expensive if it is carried out by clerical help hired for the purpose. A well-designed test should be planned so as to simplify and speed up the scoring operation. In tests for young children in the first two or three grades of school, there is not a great deal that can be done to streamline scoring procedures. Any attempt to separate the answers from the problems, so that the answers will be more convenient to score, is likely to confuse the young child and affect his score. By the upper elementary grades, however, it is practical to provide answer spaces at the side of the page, preferably the right, so that all answers appear in a column and can be scored by placing an answer key beside them.



The separate answer sheet referred to in an earlier paragraph, and also discussed and illustrated in Chapter 4, represents a further major economy in time. It completely eliminates time-consuming turning of pages by the scorer. When score is the number correct, the test can be scored by placing over the answer sheet a simple stencil with holes punched in the spaces corresponding to the right answers. There are also special types of answer sheets prepared to further simplify the scoring operation. Three main types should be noted.

1. *Carbon-Backed Answer Sheets.* (Clapp-Young, Scoreze, etc.) In these, two sheets are fastened together. On the inside certain parts of one or both sheets are covered with carbon. When the examinee marks in the answer spaces, the marks are transferred to the inside of the page by the carbon paper. The inside has the key printed upon it, in the form of boxes or circles placed opposite the correct answers. Scoring consists merely of counting the number of marks that appear in the boxes.

2. *Pin-Prick Answer Sheets.* These operate in essentially the same way, except that a pin is pushed through the answer sheet in the specified place. This technique is especially effective in the case of a multiscore test. It has been used with the *Kuder Preference Record*, where the pin is pushed through several sheets of paper, each one of which is printed with the scoring key for a different interest area. Counting the number of holes appearing within the printed circles on the different sheets gives the score for the different areas of interest without the necessity of using key or stencil.

3. *The IBM Answer Sheet and Test-Scoring Machine.* For a number of years the International Business Machines Corporation has made available a test-scoring machine that operates electrically through the conductivity of pencil marks on a special answer sheet. The sheet has 750 answer positions, which may be grouped in different ways but which most commonly represent 150 five-choice test items. The answer sheets must be rather carefully marked with a soft pencil, preferably a special one developed for the purpose, if they are to score accurately. Various other mechanical difficulties have been encountered, for example, current leakage due to a damp climate. However, when these conditions are watched for, the machine can considerably accelerate large-scale test-scoring jobs. The basic IBM machine can be bought for \$3000.00, or rented for \$50.00 per month.\* This means that the equipment must be used quite a good deal of the time if it is

\* 1960 prices.

to pay for itself. It is especially useful in organizations having a large and fairly steady flow of test scoring.

For large-scale testing programs, there are a number of agencies that maintain scoring services. These are commented on further in the discussion of school testing programs in Chapter 16.

#### FEATURES FACILITATING TEST ADMINISTRATION

In evaluating the practical usability of a test, one factor to be taken into account is the ease of administration. A test that can be handled adequately by the regular classroom teacher with no more than a session or so of special briefing is much more readily fitted into a testing program than a test requiring specially trained administrators. Several factors contribute to the ease of giving and taking a test

1. A test is easy to give if it has clear, full instructions. The instructions for the administrator should be written out substantially word for word, so that all the examiner must do is read them and follow them. Instructions for the examinee should also be complete and should provide appropriate practice exercises. The amount of practice that should be provided depends upon how novel the test task is likely to be for those being tested. Where it is a familiar type of task or a simple and straightforward instruction, no more than a single example will be needed. However, for an unusual item format or test task more practice will be desirable.

2. A test is easy to give if the number of units to be separately timed is few, and close timing is not critical. Timing a number of brief subtests to a fraction of a minute is a bothersome undertaking, and the timing is likely to be inaccurate unless a stop watch is available for each tester. Some tests have as many as eight or ten parts, each taking only 2 or 3 minutes. A test made up of three or four parts, with time limits of 5, 10, or more minutes for each, will be easier to use.

3. The layout of the test items on the page has a good deal to do with the ease of taking the test. Items in which response options are all run together on the same line, items with small or illegible pictures or diagrams, items that are crowded together, and items that run over from one page to the next all make difficulty for the examinee. Print and pictures should be large and clear. Response options should be well separated from one another. All parts of an item and all items referring to a single figure, problem, or reading passage should appear on the same page or double-page spread. Shortcomings on any of

these points represent black marks against a test as far as ease of taking it is concerned.

### FEATURES FACILITATING INTERPRETATION AND USE OF SCORES

It seems axiomatic, though the point is sometimes overlooked, that a test is given to be used. If the score is to be used, it must be interpreted and given meaning. The author and publisher of the test have the responsibility of providing the user with information that permits him to make a sound appraisal of the test in relation to his needs and to give appropriate meaning to the score of an individual. This they do primarily through the *test manual* and other collateral materials that are prepared to accompany the test. What may the test user reasonably expect to find in the manual for a test, together with its supporting materials? We have outlined below the aids we believe the test user should expect.

✓ 1. *A Statement of the Functions the Test Was Designed to Measure and of the General Procedures by Which It Was Developed.* This is the author's statement of what he considers the test to be valid for and the evidence that proper steps have been taken to achieve that validity. Particularly for achievement tests, in which we are concerned primarily with content and process validity, the author should tell us the procedures by which he arrived at his choice of content or his analysis of the functions being measured. If he is unwilling to expose his thinking to our critical scrutiny, we may perhaps be skeptical of the thoroughness or profundity of that thinking.

Procedures involve not only the rational procedures by which range of content or types of objective were selected, but also the empirical procedures by which items were tried out and screened for final inclusion in the test.

✓ 2. *Detailed Instructions for Administering the Test.* We have discussed in an earlier section the need for this aid to uniform and easy administration by the teachers or others who will have to use the test.

✓ 3. *Scoring Keys and Specific Instructions for Scoring the Test.* The problems of scoring have also been discussed, under the heading of economy. The manual and supporting materials should provide detailed instructions as to how the score is to be computed, how errors are to be treated, and how part scores are to be combined into a total score. Scoring keys and stencils should be planned to facilitate as much as possible the onerous task of scoring.

✓ 4. *Norms for Appropriate Reference Groups,* together with infor-

mation as to how they were obtained and instructions for their use. Chapter 6 was devoted to a full consideration of types of test norms and their use. It will, therefore, be sufficient at this time to point out the responsibility of the test producer to develop suitable norms for the groups with which his test is to be used. General norms are a necessity, and norms suitable for special types of communities, special occupational groups, and other more limited subgroups will add to the usefulness of a test in many cases.

✓5. *Evidence as to the Reliability of the Test.* This evidence should indicate not only the bald reliability statistics but also the operations used to obtain the reliability estimates and the descriptive and statistical characteristics of each group on which reliability data are based. If a test is available in more than one form, it is highly desirable that the producers report the correlation between the two forms. In addition to any data that were derived from a single testing. If the test yields part scores, and particularly if it is proposed that any use be made of these part scores, reliability data should be reported for the separate part scores. It is good procedure for the author to report standard errors of measurement as well as reliability coefficients. An author who indicates what the standard error of measurement is at each of a number of score levels is particularly to be commended, since this information shows over what range of scores the test maintains its accuracy.

✓6. *Evidence on the Intercorrelations of Subscores.* If the test provides several subscores, the manual should provide evidence on the intercorrelations of these. This is important in guiding the interpretation of the subscores and, particularly, in judging how much confidence to place in *differences* between the subscores. If the scores are correlated to a substantial degree, measuring much the same things, the differences between them will be largely meaningless and uninterpretable.

✓7. *Evidence on the Relationships of the Test to Other Factors.* In so far as the test is to be used as a predictive device, correlations with criterion measures constitute the essential evidence on how well it does in fact predict. Full information should be provided on the nature of the criterion variables, the group for which data are available, and the conditions under which the data were obtained. Only then can the reader fairly judge the validity of the test as a predictor.

It will often be desirable to report correlations with other measures of the same function as collateral evidence bearing on the validity of

the test. Thus, correlations with individual intelligence test score are relevant in the case of a group intelligence measure.

Finally, indications of the relationship of test score to age, sex, type of community, socio-economic level, and similar facts about the individual or the group are often helpful. They provide a basis for judging how sensitive the measure is to the background of the group and to circumstances of their life and education.

✓ 8. *Guides for Using the Test and for Interpreting Results Obtained with It.* The developers of a test presumably know how it is reasonable for the test to be used and the results from it to be evaluated. They are specialists in that test. For the test to be most useful for others, especially the teacher with limited specialized training, suggestions should be given of ways in which the test results may be used for diagnosing individual and group weaknesses, forming class groupings, organizing remedial instruction, counseling with the individual, or whatever other activities may appropriately be based on that particular type of instrument.

## SCHEDULES FOR EVALUATING A TEST

The potential user, who is trying to select the best test for a particular purpose, might welcome a standard form or procedure for evaluating the various tests that are candidates for his patronage. A standard and somewhat objective procedure for rating tests would be very attractive if an appropriate one could be devised. There have been several attempts to apply the technique of quantification to tests themselves, and score cards have been developed to be used in appraising tests.<sup>4,6,7</sup> These allocate so many points to aspects of validity, so many to factors associated with reliability, so many to ease of use and interpretation, and so forth.

One can question how useful this standard scheme of adding up points is in this situation. Certainly, if a test has low validity, no amount of elegance and polish in other respects can make it a satisfactory instrument. And the importance of different qualities for a measure varies, depending upon the purpose for which the instrument is to be used. For that reason, we are not proposing any numerical scheme for arriving at a score on each test being considered. However, a systematic outline should help in assuring that the significant factors are all taken into account and that the analysis is organized in such a way that comparison of different tests will be facilitated. The schedule given below provides such an outline. If answers are sought to all the questions raised in the outline, the potential user should have

a good basis for comparing the suitability for his needs of different available measurement devices.

An extensive and analytically critical set of criteria for an acceptable psychological test has been developed by the Committee on Test Standards of the American Psychological Association, and published by the Association. This article gives a full statement of the standards that a commercially distributed test may be expected to meet. Similar standards for educational tests have been prepared by the American Educational Research Association and the National Council on Measurements Used in Education.<sup>1</sup>

## SCHEDULE FOR EVALUATING A TEST

### GENERAL REFERENCE INFORMATION

1. Name of test.
2. Author's name (and position, if available)
3. Publisher.
4. Date of publication.
5. Cost.
6. Time for administration.

### VALIDITY

*A. Evidence from the Plan for the Test.* What were the procedures for determining the scope of the test? For determining the particular content to be covered? For determining the functions and processes to be represented? How adequate do these appear to be? How closely do the test objectives correspond to objectives that you are interested in for your school?

What provisions were made for editorial review of the test materials? How adequate do these appear?

*B. Evidence from the Test Blank Itself.* Do the test items appear appropriate for the objectives that you are trying to evaluate? Do the test items appear to be well constructed? Are they free from ambiguity? Do they have attractive wrong-answer choices?

*C. Evidence from Statistical Studies of the Test in Use.* With what concurrent measures has the test been correlated? For what sort of groups? How substantial are the correlations?

With what later criterion measures has the test been correlated? For what sorts of groups?

How does the evidence on statistical validity compare with that for other tests?

How accurate a prediction does it give of significant outside criteria? How do these results compare with those of other tests that try to measure the same trait?

*D. Evidence from Outside Authority.* What have reviewers and critics said about the validity of the test?

## RELIABILITY

*A. How Adequately Are Data Reported?* Do the authors indicate size and nature of groups for which data are reported? Do they indicate type of reliability coefficient computed? Do they give mean and standard deviation for the groups? Do they report reliabilities for single age and grade groups?

*B. What Are the Facts on Reliability?* What actual data on reliability are reported? (Indicate, as far as given, the age or grade, size of group, mean and standard deviation, procedures by which reliability was computed, and resulting values obtained.) How do the data compare with other competing tests?

## PRACTICAL CONSIDERATIONS IN ADMINISTRATION AND USE OF TEST

*A. Factors in Administration*

1. Adequacy of manual.
2. Complexity of procedures.
  - a. Complexity of process required of students.
  - b. Adequacy of instructions and practice exercises.
  - c. Complexity of process required of examiner. Timing, giving instructions, and interpreting responses of subjects examined.
3. Time requirements.
4. Legibility, attractiveness, and convenience of format.

*B. Factors in Scoring*

1. Time required (i.e., form of answer, type of key, etc.).
2. Special skills required (subjective scoring and qualitative interpretation).

*C. Factors in Interpretation*

1. Type of norms. Appropriateness to uses, completeness, representativeness of sample. How readily may raw scores be converted into derived scores?
2. Aids to interpretation provided by manual.

*D. Factors in Continued Use*

1. Are there comparable forms? How many? How well is comparability established?
2. Cost. Does this permit routine continued use? Can blanks be used a number of times?

## ✓ SUMMARY STATEMENT

We have discussed the requirements of a good test under the headings of validity, reliability, and practicality. A test is valid in so far as it measures the qualities we wish to measure. It is reliable in so far

as it measures with precision. It is practical in so far as it is economical of time and money and simple to give and interpret.

The crucial requirement for a test is validity. In some tests, especially achievement tests, we may have to judge how well the test *represents* the content and processes we wish to measure. For other tests, especially aptitude tests, we may evaluate how well the test *predicts* some measure that serves as a later criterion of success. In still others, where we are interested in the test as *describing* some trait or aspect of the individual, appraisal of validity is more complex. A "theory" of the trait or construct must be developed, and the test is evaluated by how well it fits into the pattern of relationships that would be predicted from this theory.

There are several different procedures available for obtaining estimates of the *reliability* or precision of a measure. The most rigorous procedure is to administer two *equivalent forms* of the test on two separate occasions. The correlation between the two forms provides a reliability coefficient that tells how closely individuals maintain their position in the group from one testing to the other. Less exacting procedures include (1) repetition of the same test and (2) extracting two scores from a single test, usually by scoring odd and even items separately. Reliability estimates based on these last procedures are less satisfactory and should usually be discounted somewhat.

The value obtained for the reliability coefficient will depend on the range and level of ability in the group tested and the length of the test, as well as upon the particular procedure used for estimation. It is particularly necessary to discount a coefficient based on the pooling of several grades.

To describe the accuracy of an individual's score, the *standard error of measurement* is often preferable to the reliability coefficient. It tells the variation to be expected if we were able to make repeated measurements of a particular individual. This variation must always be borne in mind when interpreting the score an individual receives.

Practicality is a function of economy, ease of administration, and readiness of interpretation. Economy is affected by initial cost, by the possibility of reusing materials, and by time required for scoring and analyzing the results. Ease of administration results from full directions, simple procedures for the examinee, and an objective record of performance. Readiness of interpretation is facilitated by good norms and by a full guide of suggestions for interpretation.

The potential test user should examine the tests from among which he must choose in the light of the above criteria and pick the one that best fits his needs.



## REFERENCES

1. American Educational Research Association and National Council on Measurements Used in Education, Committee on Test Standards, *Technical recommendations for achievement tests*, Washington, D. C., National Education Association, 1955.
2. American Psychological Association, Committee on Test Standards, Technical recommendations for psychological tests and diagnostic techniques, *Psychol. Bull.*, 51, No. 2, Pt. 2, 1954.
3. Buhler, R. A., Flicker fusion threshold and anxiety level, unpublished doctor's dissertation, Columbia University, 1953.
4. Cole, R. D., and F. von Borghersrode, A scale for rating standardized tests, *Sch. of Educ. Rec. of Univ. of North Dakota*, 14, 1928 (Oct.), 11-15.
5. McClelland, David, John W. Atkinson, Russell A. Clark, and Edgar L. Lowell, *The Achievement Motive*, New York, Appleton-Century-Crofts, 1953.
6. Otis, A. S., *Scale for rating tests*, Yonkers, N. Y., World Book, 1926.
7. Rinsland, H. D., Form for briefing and evaluating standardized tests. *J. educ. Res.*, 1949, 42, 371-375.
8. Terman, L. M., and Maud A. Merrill, *Measuring intelligence*, Cambridge, Mass., Houghton Mifflin, 1937.

## SUGGESTED ADDITIONAL READING

- American Educational Research Association and National Committee on Measurements Used in Education, *Technical recommendations for achievement tests*, Washington, D. C., National Education Association, 1955.
- American Psychological Association, Committee on Test Standards, Technical recommendations for psychological tests and diagnostic techniques. *Psychol. Bull.*, 51, No. 2, 1954, Pt. 2.
- Bennett, George K., Harold G. Seashore, and Alexander G. Wesman, *Differential Aptitude Tests manual*, New York, Psychological Corp., 1959, Chapters 4 and 5.
- Cronbach, Lee J., *Essentials of psychological testing*, 2nd ed., New York, Harper, 1960, Chapters 5 and 6.
- Cureton, Edward E., Validity, Chapter 16 in E. F. Lindquist, Editor, *Educational measurement*, Washington, D. C., American Council on Education, 1951.
- Doppelt, Jerome E., How accurate is a test score? *Test Service Bulletin No. 50*, New York, Psychological Corp., 1956.
- Harris, Chester W., Editor, *Encyclopedia of educational research*, 3rd ed., New York, Macmillan, 1960, pp. 1038-1047, 1551-1554.
- Thorndike, Robert L., Reliability, Chapter 15 in E. F. Lindquist, Editor, *Educational measurement*, Washington, D. C., American Council on Education, 1951.
- Wesman, Alexander G., Expectancy tables—a way of interpreting test validity, *Test Service Bulletin No. 38*, Psychological Corp., 1949, 1-5.

## QUESTIONS FOR DISCUSSION

1. If the College Entrance Examination Board were developing a general survey test in science for high-school seniors, what might they do to establish the validity of the test?

2. What type of validity is indicated by each of the following statements which might be found in a test manual?

- Scores on Personality Test X correlated  $+ .43$  with teachers' ratings of adjustment.
- The objectives to be appraised by Reading Test Y were rated for importance by 150 classroom teachers.
- Scores on Clerical Aptitude Test Z correlated  $+ .57$  with supervisors' ratings after 6 months on the job.
- Intelligence Test W gives scores that correlate  $+ .69$  with *Stanford-Binet* IQ.
- Achievement Battery V is based on an analysis of 50 widely used texts and 100 courses of study from all parts of the U. S.

3. Comment on the statement "The classroom teacher is the only one who can judge the validity of a standardized achievement test for his class."

4. Look at the manuals of two or three tests of different types. What evidence on validity is presented? How adequate is it for each test?

5. Using Table 7.3 on p. 171, determine what per cent of those selected would be above average on the job if a selection procedure with a validity of  $.40$  were used and only the top quarter were accepted for the job. What per cent would be above average if the top three-quarters were selected? What would the two per cents be if the validity were  $.50$ ? What does a comparison of the four percentages bring out?

6. Air Force personnel psychologists are doing research on the selection of jet-engine mechanics. What might they use as criterion measures of success as a mechanic? What are the advantages and limitations of each possible measure?

7. What advantages and disadvantages do school grades have as criterion measures?

8. A test manual contains the following statement: "The validity of test X is shown by the fact that it correlates  $.80$  with the *Stanford-Binet*." What additional information is needed to evaluate this assertion?

9. Look at the evidence presented on reliability in the manuals of two or three tests. How adequate is it? What are its shortcomings?

10. The manual for test T presents reliability data based on (a) retesting with the same test form a week later, (b) correlating odd with even items, and (c) correlating form A with form B, the two forms being given a week apart. Which procedure may be expected to yield the lowest coefficient? Why? Which to yield the most useful estimate of reliability? Why?

11. A student has been given the *Stanford-Binet Intelligence Test* four different times during his school career, and his cumulative record card shows the following IQ's: 98, 107, 101, and 95. What significance should be attached to the fluctuations in IQ?

12. A school plans to give form A of a reading test in October and form B in May, in order to study individual differences in improvement during the year. The reliability of each form of the test is known to be about .85 for a grade group. The correlation between the two forms turned out to be .80. How much confidence can be placed in the "gain" scores?

13. You are considering three reading tests for use in your school. As far as you can judge, the three are equally valid. The reliability of each is reported to be .90. What else would you need to know to make a choice among the tests?

14. Examine several tests of intelligence or of achievement that would be suitable for a class you are teaching or might teach. Write an evaluation of one of these tests, following the guide on pp. 201-202.

## Chapter 8



# Where to Find Information about Specific Tests

### THE NATURE OF THE PROBLEM

The production of educational and psychological tests has been going on for only half a century, but during that time literally thousands of different tests have been produced. In a comprehensive bibliography which covered up to about 1945, Hildreth <sup>8,10</sup> included entries for 5294 different tests. The number could probably now be increased by at least another thousand. Of course, many of the earlier published tests are obsolete now, or only of historical interest, but the number of currently available tests is still very great.

Not only is the total number of tests great. So also is the variety. Tests vary widely in testing procedures, in content, and in group for which designed. There are paper-and-pencil tests, individual performance tests, rating scales, self-rating procedures, observational procedures, and projective techniques. There are measures of attitude, of interest, of temperament, of personal adjustment, of intellect, of special aptitudes, and of all aspects of school achievement. There are tests designed for infants, for preschool children, for school children and adolescents, and for adults.

No one book can hope to introduce a student to even a representative sampling of tests of all types, covering all sorts of content for all age levels. The following chapters will introduce some of the most important and most widely known tests, discussing them as examples of many others. But this book cannot give a complete treatment of any particular age group or subject area, and there are so many special situations in which a reader may be interested or for which he may need a test that the tests discussed here may include not even one that fits his particular need.

Since it is impossible to list and evaluate all or even most of the tests that might be of concern to an audience with varied interests, we

shall approach the problem at a different level. We shall try to guide the reader to sources in which he can find the available tests listed, and in some cases evaluated, and we will try to guide the reader in evaluating the tests he locates. The present chapter discusses resource materials for finding tests and for finding out about them. Chapter 7 has given an orientation in the factors to be considered in evaluating the suitability of a particular test for a particular purpose.

The knowledge of where to go to find out about tests of a particular type and how to evaluate one when found is probably more important than predigested information about a particular test. Tests change and the purposes of the test user change. It is impossible to anticipate what type test will be required for some future need. The important thing is to know how to go about finding the tests available for that need when it arises and how to evaluate their relative merits.

There are several different types of questions about a test or an area of measurement for which one may seek answers. Some of the types of questions are:

1. What tests have been developed that might serve my present need or purpose?
2. What are the *new* tests in my field of interest?
3. What is test A, of which I have heard, like? For what groups and purposes was it designed? Who made it? How long does it take and how much does it cost? What skills are needed to give and use it?
4. What do specialists in the field of measurement have to say about test A? How do they evaluate it, in comparison with competing techniques?
5. What basic factual material do we have on test A? What are its statistical attributes? What are its relationships to other measures?
6. What research has been done studying or using test A?

Let us see what materials are available to us as we try to answer questions such as these. These resources include (1) text and reference books in special areas of testing, (2) the *Mental Measurements Yearbooks*, (3) test reviews in professional journals, (4) publishers' test catalogues, (5) each test itself together with its accompanying manual, (6) articles in professional journals reviewing a broad field of testing, (7) comprehensive bibliographies of tests and the testing literature, and (8) educational and psychological abstract and index series. These will be considered in turn, the most useful items will be identified, and the information to be obtained from each type of source will be indicated.

## TEXT AND REFERENCE BOOKS IN SPECIAL AREAS

There are a number of text and reference books covering more specialized areas of testing. When the scope is limited to include only elementary-school tests, tests for diagnosis of individual maladjustment, or tests for vocational placement, it becomes possible to cover the field in more detail. A book dealing with tests of a particular type provides a good general introduction to the materials of the field. Such a book usually acquaints the reader with a representative selection of established tests in the area—those which the author considers worthy of mention. In addition, some evaluation of each test is usually given, indicating the purposes for which it may well be used, and what the writer considers to be its strengths, weaknesses, and distinctive characteristics. The book will usually also contain some discussion of the problems of testing in the field it covers, apart from discussion of specific tests.

It is not possible to consider all the books that might prove useful to some reader. However, a number of them have been listed below with brief annotations. The titles have been chosen in terms of their recency and the quality of their treatment. In addition, an attempt has been made to get books that represent a wide range of specialized interests. The annotations are designed to bring out the distinctive quality of each book.

- Allen, Robert M., *Personality assessment procedures*, New York, Harper, 1958. Surveys methods and techniques for evaluating personality, and is a source book of tests and instruments used to assess personality. For each test, the purpose, reliability, validity, and standardization procedures are given.
- Arny, Clara Brown, *Evaluation in home economics*, New York, Appleton-Century-Crofts, 1953. Although the examples given in the first half are related to home economics, the excellent discussion of purposes and methods of evaluating student progress are applicable to any class. Commercially published standardized tests, check lists, and rating scales are described and uses indicated in an appendix.
- Bauman, Mary K., *A manual of norms for tests used in counseling blind persons*, AFB Publications, Research Series, No. 6, New York, American Foundation for the Blind, 1958. Information is given on tests that have been adapted or developed for use with the adult blind. Also includes bibliography of source material on tests, testing, and test interpretation for adult blind.
- Blair, Glenn M., *Diagnostic and remedial teaching: a guide to practice in elementary and secondary schools*, rev. ed., New York, Macmillan, 1956. Gives selected lists of tests which Blair judges to be of value for diag-

nosis of difficulties in the basic skills. Comments on each test are given indicating strengths and weaknesses and suggested ways of using.

Bond, Guy L., and Eva Bond Wagner, *Teaching the child to read*, 3rd ed., New York, Macmillan, 1960. Appendix contains information on reading readiness tests, diagnostic and survey tests of reading, group and individual intelligence tests.

Clarke, H. Harrison, *Application of measurement to health and physical education*, 3rd ed., Englewood Cliffs, N. J., Prentice-Hall, 1959. Describes a variety of performance tests for physical fitness and skills, paper-and-pencil tests for knowledge of sports techniques and health education, and rating scale techniques. Emphasizes use of tests and need for planning an efficient program for evaluating students in physical education.

Froehlich, Clifford P., and Kenneth B. Hoyt, *Guidance testing*, 3rd ed., Chicago, Science Research Associates, 1959. Chapters in book are devoted to scholastic ability, multifactor aptitude batteries, single aptitude tests, achievement tests, interest inventories, and personal and social adjustment inventories. At the end of each chapter is a list of tests judged by Froehlich and Hoyt to be the most useful and best in the area.

Hardaway, Mathilde, and Thomas Maier, *Tests and measurements in business education*, 2nd ed., Cincinnati, South-Western Publishing Co., 1952. Provides lists of achievement and prognostic tests available in business education.

Super, Donald E., and John O. Crites, *Appraising vocational fitness*, New York, Harper, in press. Reports on selected tests in a wide variety of fields that may be used in educational or vocational guidance.

One limitation of books, such as those just annotated, becomes apparent from an examination of the publication dates. At the time that these were selected (1960), each was judged to be the most recent good book in its field and yet some were already eight years old. When one adds to this the time that has elapsed in the preparation and printing of the book, it is easy to see that a book reviewing a field cannot be relied upon for current materials. The typical textbook gives information about well-established and accepted tests, but recently published devices or techniques that are still in the experimental stages are not likely to be represented. There is a lag of several years between production of a device and the reporting of it in books reviewing an area of testing.

Another feature of most books surveying a field, which may be in some cases an advantage and in others a disadvantage, is that they are selective. They must be. The author cannot discuss everything, so he must pick the items he wishes to present. He selects for discussion the tests which he considers valuable. In so far as his judgment is sound, he does a real service to the novice in the field, who is thus led directly to the more important and valuable material. However, this means that the reader cannot expect to use a textbook as a source to lead him

to all the tests in an area and permit him to compare them. For a full listing of the tests of any particular type he will have to look elsewhere.

## THE MENTAL MEASUREMENTS YEARBOOKS

Probably the most useful single reference source for the person needing to make choices and plan programs in the field of testing is the series of *Mental Measurements Yearbooks* prepared by Buros.<sup>3</sup> 4,5,6,7 Five *Yearbooks* have now been published, and they were preceded by two more modest volumes of the same type. The *Yearbooks* undertake to provide a listing and one or more frank and critical reviews of each new standardized test that is published.

A large panel of reviewers has cooperated in the preparation of these volumes, each reviewer evaluating two or three tests in an area in which he is presumed to be competent. The tests of more general interest are appraised by two and sometimes even more reviewers. The reviews are fairly full, pointing out strengths and weaknesses of a test, comparing it with others in the field, and indicating the purposes for which the reviewer considers it useful.

In addition to reviews of tests, the *Yearbooks* also include the factual items about each test that a potential user is likely to need—such items as author, publisher, publication date, cost, time to administer, grades for which suitable, and number of forms available. Finally, for each test the *Yearbooks* give a bibliography of books and articles that have appeared dealing with that particular test. These bibliographies are quite extensive, amounting in the case of one test to 2297 titles.

The *Yearbooks* have two other features that add to their value to the test user. One is a section on books and monographs related to measurement problems. This section undertakes to list all the significant books on measurement for the period covered and in addition gives excerpts from the reviews of these books that have appeared in psychological and educational journals. The bibliography and reviews provide a guide to, and evaluation of, publications in the field.

Also valuable is a very complete index and directory section. This includes (1) a directory and index of the publishers of the tests and of the books on measurement reviewed in the volume, (2) a directory and index of the periodicals that have included reviews of tests or books on testing, (3) an index of titles of books and tests, (4) an index of names occurring in any connection, and (5) a classified index of tests organized by content or type. These indices make it possible to locate any test or type of test, to locate the complete original of any



excerpted test review, and to get in touch with the publisher of any test.

When a question arises about a test or a type of test, the *Mental Measurements Yearbooks* are the volumes for which one reaches almost automatically. They are a "must" for any individual or any office that must answer frequent questions about tests or testing.

The *Yearbooks* are not too convenient to use if one wishes to cover early as well as current tests in a particular area. At the present writing, there are five of them, published in 1938, 1941, 1949, 1953, and 1959. To cover the tests in any field, the reader must search all five volumes and may in fact need to go back to antecedent publications.<sup>1,2</sup>

A new test is ordinarily reviewed in the first *Yearbook* that came out after it was published, and reviews may sometimes also appear in subsequent volumes. Space limitations did not permit review in the 1938 *Yearbook* of all the older tests that were thought to merit review, and reviews of some of these first appeared in later volumes. Even the set of volumes taken together does not undertake to be *exhaustive* in its coverage of tests of a given type. However, if he brings together the material in the complete series, the reader will probably find an appraisal of any test that he is likely to consider using, published up to the time that planning for the last *Yearbook* was completed. The first two *Yearbooks* cover tests up to about 1939; the third covers the period from 1940 through 1947; the fourth deals with material from 1948 through 1951; and the fifth brings us up to 1958.

### JOURNAL TEST REVIEWS

We still face the problem of getting information on the *latest* tests and testing developments. One way of keeping up with important new tests is through reviews in professional journals. Tests of interest to the psychologist and the counselor have been reviewed for a number of years in the *Journal of Consulting Psychology* and the *Journal of Counseling Psychology*. In late 1959, a test review section, called "Testing the Test," was initiated in the *Personnel and Guidance Journal*. These sources should keep the test user up to date on the most significant new psychological tests within a year or so of their appearance.

### TEST PUBLISHERS

The most up-to-date information on what tests are available is probably to be obtained from the test publishers themselves, either through correspondence or through their catalogues. There are many pub-

lishers, too many to list here, so that gathering information from all of them would be quite an undertaking. However, the number who publish *extensively* in the testing field is a good deal more limited. A number of the most important publishers are listed in Appendix IV together with their addresses and some indication of the types of material and the services they supply.

The limitations of a test publisher as an entirely unbiased source of information on the *values and limitations* of his own publications are, of course, obvious. Reversing Marc Antony, we may say he comes to praise his tests, not to bury them. However, as a source of information about, rather than evaluation of his tests, he can be very helpful. In Chapter 7 we have considered how the potential user may go about appraising a new test for himself in the light of the information he can get from the test producer and from other sources.

## TEST AND MANUAL

The individual who is seriously considering using a particular test will certainly need to examine the test itself and the manual the publisher has prepared to go with it. Each publisher's catalogue will indicate the price for which a specimen set of each test may be obtained. The specimen set contains a copy of the test itself, the instructions for administering and scoring, and part or all of the supplementary materials available to the user to help in interpreting the test.

The amount of supplementary materials included in a specimen set varies from one publisher to another. The potential user can legitimately expect the publisher to include materials in a specimen set that will provide all the information he needs in order to arrive at a decision as to the suitability of the test for his purposes. He should be skeptical of any test for which the information supplied him is incomplete. The individual who wishes to examine a number of different tests without buying specimen sets of each may be able to find a test file in the library or the guidance department of his local university.

To obtain specimen sets of tests, the applicant must ordinarily present some sort of credentials. A letter on the official letterhead of his school or institution will often suffice. A note from the university where he is studying may serve the function. The limitations that publishers place upon the distribution of their materials depend upon the nature of the materials. They will often refuse to distribute tests that require special skills to administer and interpret unless the ap-

plicant can give evidence that he has the training and skills that qualify him to use the materials.

A detailed examination of the test itself will provide the potential user with a basis for judging how well the content of the test and the form of test exercises correspond to the objectives and functions he wishes to measure. The accompanying material, which we have collectively called the test manual, is a very important part of any test. It varies enormously in quality and comprehensiveness from one test to another. In some of the better current tests, this collateral material becomes almost a book. It provides a great variety of important information to help in using and interpreting a test. We have indicated in Chapter 7 (pp. 202-203) the types of information a test user has a right to expect to find in the test manual. A manual that provides all this information becomes a very important source for information about the test.

Manuals differ greatly not only in comprehensiveness but also in impartiality and integrity. Probably no test manual is entirely free of a promotional element. However, sometimes the manual becomes to a very large extent a promotional device focused on increasing the sales of the test. The potential user must always be aware of this aspect of the manual and must endeavor to discount appropriately claims made for the test. There often appears to be an inverse relationship between the grandeur of the claims that are made and the evidence on which they are based. The reader will do well to keep his attention focused on the evidence presented in the manual, to view claims in the light of this evidence, and to be extremely suspicious of the test whose manual makes sweeping claims but presents very little data.

### JOURNAL REVIEW ARTICLES

It is sometimes useful to refer to summary articles covering recent developments in tests and testing. The most regular of these in recent years has been the triennial summary in the *Review of Educational Research*. This journal undertakes to summarize research in a number of different sectors of education. Its publication schedule is arranged so that a given area is treated every 3 years. Material on tests and measurements was reviewed in the February, 1959, issue, which was devoted to educational and psychological testing. Similar reviews appeared in 1956, 1953, and every third year back to 1932. Because of the volume of material to be covered, these reviews are very condensed, but they do introduce the reader to new tests and

testing research and provide him with a bibliography of original references to which he can go for a fuller report on any topic in which he is interested.

Since 1950, the *Annual Review of Psychology* has provided a yearly review and bibliography on selected psychological topics. Chapter headings such as "Individual Differences" and "Theory and Techniques of Assessment" suggest sources for material of possible interest to the psychological tester.

Gray has for many years prepared an annotated bibliography on reading, which has appeared in recent years in the *Journal of Educational Research*. This deals with reading tests—as well as with other reading problems.

### COMPREHENSIVE BIBLIOGRAPHIES ON TESTS AND TESTING

There have been, from time to time, comprehensive bibliographies on tests and testing. However, the most complete of these, by Hildreth<sup>9,10</sup> is badly out of date, covering material only up to 1945, and is of interest primarily to a person having a historical interest in tests of a given type. Its coverage up to the time of its publication was quite complete. The bibliography merely lists and gives a reference source for each test, and provides no further information about it.

One fairly extensive bibliography dealing with the technical aspects of testing (i.e., such issues as the appraisal of reliability, item analysis techniques, etc.) has been prepared by Goheen and Kavruck.<sup>6</sup> This source provides over 2500 references covering the period from 1929 to 1949.

These extensive bibliographies will be useful primarily to the person who wants to dig fairly deeply into tests of some particular type, or into some technical testing problem.

### ABSTRACTS AND INDICES

Two final sources that must be brought to the attention of the serious student are the *Psychological Abstracts* and the *Education Index*. These are basic bibliographic sources in the fields of psychology and education respectively. Each undertakes to provide a complete listing of current publications in its respective field. The field for the *Psychological Abstracts* is rather more narrowly defined, being restricted to scientific and technical publications in psychology. Each publication is represented not merely by title but also by an abstract indicat-

ing the nature of the report and the major findings. An annual subject index and author index aid in locating desired material. The *Psychological Abstracts* provides a monthly listing of new tests. This appears in the "General" section at the beginning of each issue under the heading New Tests. The *Abstracts* also covers the literature of research using tests and of findings with respect to them.

The *Education Index* covers a considerably wider range of material, since it deals with the whole broad area of education and includes popular and professional materials as well as those of a more technical and scientific nature. It gives references only, providing no information about the nature and content of the item. Material is topically organized, and the user who looks under such topics as ability tests, educational measurement, mental tests, or personality tests will find most of the material relating to measurement in education.

The joint use of the *Psychological Abstracts* and the *Education Index*, supplemented by the other sources discussed previously, should enable the student who wishes to dig to the roots of a measurement problem to locate the bulk of the work that has been done on that problem.

## SUMMARY STATEMENT

At the beginning of this chapter a number of questions were suggested to which a test user might wish answers. The important sources of information about tests and testing have now been discussed. By way of summary, we may try to relate the sources to the questions. An attempt has been made to do this in Fig. 8.1. At the top of this chart are listed various questions one might raise about a test, type of test, or testing problem. On the side are listed the most important types of source material referred to in this chapter. In each cell is a symbol to represent the extent to which the source should help in answering the question. The symbol \*\* is used to designate one of the sources that would probably be *most* helpful and to which one would turn first. Sources marked \* are ones that would also be expected to contribute to the needed answer. Sources marked ? are ones that might perhaps provide some useful information. Where there is *no* entry at all, the source is not likely to be helpful in that connection. A critical study of this table, with analysis of the reasons for the various entries, should leave the reader well prepared to go out and get for himself the information he needs in order to select a test or as background for a specific testing problem.

7. Buros, O. K., *The fifth mental measurements yearbook*, Highland Park, N. J., Gryphon Press, 1959.
8. Goheen, Howard W., and Samuel Kavruck, *Test construction, mental test theory, and statistics*, Washington, D. C., U. S. Government Printing Office, 1950.
9. Hildreth, Gertrude H., *A bibliography of mental tests and rating scales*, New York, Psychological Corp., 1939.
10. Hildreth, Gertrude H., *A bibliography of mental tests and rating scales. 1945 supplement*, New York, Psychological Corp., 1946.

## QUESTIONS FOR DISCUSSION

1. Using the sources indicated in the text, prepare as complete a list as you can of currently available standardized tests for a specific grade and purpose (i.e., tests in first-year Spanish, reading readiness tests, tests in American history for the twelfth grade, etc.).

2. Using the *Mental Measurements Yearbooks*, find out what reviewers think of a particular test that you are interested in.

3. Using the *Fifth Mental Measurements Yearbook*, find out what reviewers have to say about one of the following titles that interests you:

Doll, E. A., *The Measurement of Social Competence*.

Eysenck, H. J., *The Scientific Study of Personality*.

Remmers, H. H., *Introduction to Opinion and Attitude Measurement*.

Sarason, S. B., *The Clinical Interaction: With Special Reference to Rorschach*.

Strong, E. K., Jr., *Vocational Interests 18 Years After College*.

4. To what sources would you go to try to answer each of the following questions? To which would you go first? What would you expect to get from each?

- a. What test should I use to study the progress of two class groups in beginning French?
- b. What kinds of norms are available for the *Stanford Achievement Tests*?
- c. Is the *Rorschach Test* of any value as a predictor of academic success in college?
- d. Has a new revision of the *Wechsler Adult Intelligence Scale* been published yet?
- e. What intelligence tests have been developed for use with the blind?
- f. What are the significant differences between the *Metropolitan Achievement Tests* and the *California Achievement Tests*?
- g. How much does the *Otis Quick-Scoring Intelligence Test—Beta*, cost?
- h. What do testing people think of the *Brainard Occupational Preference Inventory*?

5. Look at two or three publishers' catalogues. Compare the announcements of tests of the same type. How adequate is the information that is provided? How objective is the presentation of the tests' values and limitations?

## Chapter 9



# Standardized Tests of Intelligence or Scholastic Aptitude

### ACHIEVEMENT AND APTITUDE

Ability tests are designed to appraise what an individual *can* do under favorable conditions when he is trying to do his best. All any ability test measures is performance at the time of testing. From this performance we may hope to make one or more of a variety of different inferences. We may want to infer how effective a program of school instruction has been in teaching new knowledges or skills, i.e., how much progress the pupils have made in some kind of achievement. We may want to infer how well each individual will do in learning some new task, i.e., a prognosis of future achievement. We may want to make inferences about the organization or structure of human abilities, i.e., what goes with what. We may hope to unravel the causal factors in individual abilities or disabilities, i.e., why the individual fails or succeeds with a particular task. All these are different sorts of *inferences*. The basic evidence in every case is performance on a set of test tasks.

Performances are tied with varying degrees of closeness to specific, organized instruction. At one end of the scale are those knowledges and skills that are the direct outcome of organized teaching, usually in schools but sometimes on the job. To decipher the meaning of: *Arma virumque cano* or of

are accomplishments that will be developed almost exclusively in a high-school course in Latin, on the one hand, or Gregg shorthand, on the other. Even the ablest individual with a wealth of general

life experiences is unlikely to acquire abilities such as these unless they have been specifically taught. We frequently want to measure the extent to which abilities such as these, dependent directly upon formal instruction, have been acquired. Tests thus tied to instruction and concerned with evaluation of past progress are spoken of as *achievement tests* or *proficiency tests*. We shall consider them in some detail in Chapter 11.

At the other end of the scale are abilities that are developed through the general experiences of life, quite apart from any formal instruction. Consider the two pictures in Fig. 9.1A and B. Suppose we were to ask a child, concerning each of them: What is wrong with this picture? What is silly about it? As we went up the age range, we would find more and more children who could give us a satisfactory answer. But probably no child would have been specifically taught in school that shadows extend away from the sun or that in a wind flags and smoke will be blowing in the same direction. The background to apprehend the absurdity in these situations and the ability to isolate the critical elements in the pictures come with maturity from the general experiences of growing up in our society.

It should be emphasized that any performance depends in some degree upon experience. A child from a culture that had provided no experience with books and pictures would be less likely to suc-

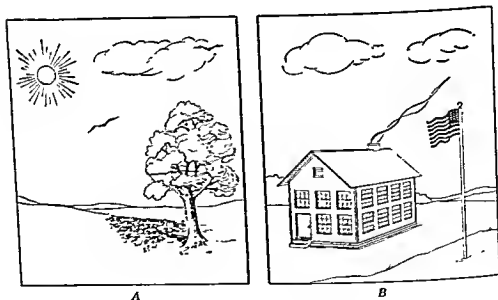


Fig. 9.1A and B. Picture-absurdity-test items.



ceed with the tasks of Fig. 9.1 because he had never learned to interpret a picture as corresponding to real things in a real world. A child who had had no experience with chimneys or with flags would be severely handicapped on picture *B* since he would not be able to interpret the picture or know how these things should behave. These two absurdities items assume (1) a general familiarity with pictures and the representation of things by pictures and (2) experience with trees, shadows, houses, flags, smoke, and wind. Any child in a normal American environment will have had these experiences in abundance. For him, therefore, the test provides a measure of perception, analysis, and understanding of his environment. Differences between individuals in performance on these tasks may then reasonably be expected to reflect fairly basic differences in certain aspects of intellectual ability.

The examples we have given have illustrated two points quite far apart on the scale ranging from "directly taught" to "acquired entirely from general life experiences." Many abilities fall at intermediate points along this scale. The meanings of words, for example, are taught in school in connection with almost every segment of the school program. But a very large part of our stock of word meanings is picked up in the reading and listening done out of school as an incidental by-product of just living in our society. Again, reading is usually first learned in school, but a large part of the growth in fluency of reading and depth of understanding of printed matter comes from out-of-school reading and from the general acquiring of experience and maturity as a part of growing up. There is no clear boundary line marking off the ability that is a school achievement from the one that is not.

Psychologists and educators are interested in measuring the underlying aptitudes of human beings. The interest is sometimes in using these aptitude measures to predict later achievements. It is sometimes in studying the aptitudes for their own sake. But the concept of aptitude is a tricky one. Aptitude implies some natural or innate capacity for a particular type of performance—scholastic aptitude, mechanical aptitude, or artistic aptitude. But all we can observe is performance on a set of tasks. As stated above, this performance inevitably depends in some measure upon the experiences that the individual has had. If we want to get at basic individual differences in capacity to do a certain type of task, our only hope is to seek for test items based on experiences so common and general in our culture that almost every person will have had the requisite experiences. We

must build upon the common core of experience available to all. This is what aptitude tests aspire to do. They try to base their items upon experiences, mostly out of school but overlapping to some extent those provided in school, that are uniformly provided for individuals growing up in our society. They use these present abilities, based inevitably on a variety of past learnings, as indicators of what the individual can learn to do in the future.

The difference between *aptitude* measures and *achievement* measures is, then, one of degree and emphasis. Any test of ability is to some extent an aptitude test and to some extent an achievement test. The difference between the two designations is perhaps as much in the type of inference that we want to make as in the specific content or the "innateness" of the measure. A test can be thought of as an achievement test when we wish to draw conclusions about past progress and as an aptitude test when we wish to estimate future potentialities. The remainder of the present chapter will be devoted to tests of general intellectual ability, or scholastic aptitude. Chapter 10 will be concerned with other types of special abilities, and Chapter 11 will be devoted to standardized tests of educational achievement.

### TASKS USED TO MEASURE ABSTRACT INTELLIGENCE

Much of the research and development of aptitude measures has been devoted to devising and studying tests of "general intelligence," familiarly known as "IQ tests." General intelligence, in this context, has typically meant abstract intelligence—the ability to see relations in, make generalizations from, and relate and organize ideas represented in symbolic form. What general intelligence has meant to those who have tried to test it can be seen from the types of tasks they have used. Examples of a number of the common types of tasks are given below. The keyed answers for multiple-choice items are underlined.

#### VOCABULARY

A word meaning nearly the same as *robust* is

- A. cheerful.    B. strong.    C. fat.    D. small.    E. wealthy.

#### VERBAL ANALOGIES

Branch is to tree as brook is to

- A. water.    B. root.    C. bank.    D. river.    E. habble.

## SENTENCE COMPLETION

The sun rises in the \_\_\_\_ and sets in the west.

- A. summer.    B. morning    C. east    D. end    E. sky

## ARITHMETIC REASONING

A boy bought candy bars at 90 cents for a box of 24 and sold them at 5 cents each. How much did he make on each bar?

- A. 30 cents.    B. 3½ cents    C. 1½ cents    D. 1 cent  
E. None of these

## NUMBER SERIES

What number should come next to continue the series 1, 2, 4, 7, 11?

- A. 14    B. 15    C. 16    D. 18    E. 22

## FIGURE ANALOGIES

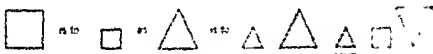


Fig. 910

## CLASSIFICATION

Look at the three words on the left. Which word on the right belongs with these three?

Doctor    Lawyer    Engineer

Farmer    Architect    Mechanic  
Carpenter    Doctor

## "MULTIMENTAL"

Which one of the figures does not belong with the other four?

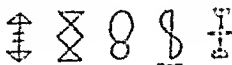


Fig. 912

## PICTURE ARRANGEMENT

The pictures below tell a story. Which picture comes first in the story?



Fig. 9.1E.

## COMPREHENSION (COMMON SENSE)

What is the thing to do if you bump into someone and hurt him?

## SIMILARITIES

In what way are wool and cotton alike?

## INFORMATION

What month in the year has the fewest days?

## DIGIT SPAN

"I will say some numbers. Listen carefully, and when I am through repeat exactly what I said. Listen—

3 8 7 1 5

Now repeat what I said."

## DIGIT SYMBOL SUBSTITUTION

This is a code test. Each figure stands for a particular number. You are to put the right numbers in the boxes as fast as you can.

Code

△	○	▱	×	8
1	2	3	4	5

Test

△	8	×	○	△	▱	8	×	△	8	etc.

Fig. 9.1F.

## OBJECT ASSEMBLY

These pieces, if put together correctly, will make a boy. Go ahead and put them together.

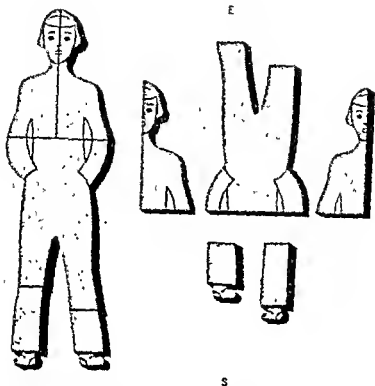


Fig. 9.1G. Object Assembly Test item from Wechsler Intelligence Scale for Children. (Reproduced by permission of the Psychological Corporation)

## GROUP INTELLIGENCE TESTS

Most of the intelligence testing carried on in this country is done with group tests. These are paper-and-pencil tests much like the objective type of school examination. They usually consist of 75 to 100 multiple-choice items of the types illustrated in the previous section. Ordinarily, the examinee must read the problem to himself, must work along and do the tasks one after another, and must do as many as he can within a fixed time limit. However, some group tests call for oral instructions from the examiner, and some are paced by the rate at which the examiner presents the test tasks.

Some group intelligence tests (e.g., *California*, *Kuhlmann-Anderson*, *Large-Thorndike*, *Pintner*) are made up of several separately timed subtests, in each of which all the items follow the same pattern; i.e.,

all are vocabulary items or all are number series items. Others (e.g., *Henmon-Nelson, Otis*) have the different types of items mixed in together, a vocabulary item being followed by a number-series item, that by a figure-analogies item, etc. The cycle of different types of items is repeated, the items gradually becoming more difficult. This type of test is called a "spiral omnibus" test because of the cyclical pattern.

The typical group test is designed to cover a range of three or four school grades, i.e., 4 to 6, 7 to 9, 10 to 13. Tests for elementary-school children usually call for responses marked in the test booklet itself, but many of the tests for older groups use separate answer sheets that can be machine scored.

There are a number of different series of group tests on the market that are quite satisfactory to use. The number is too great to permit discussion of each one here. Several are listed, together with annotations describing and providing some evaluation of each, in Appendix III.

In the remainder of this chapter we will first describe the two individual tests, i.e., tests given to one examinee at a time in a face-to-face setting, that are currently most widely used in the United States. These are the *Stanford-Binet Intelligence Scale* and the *Wechsler Adult Intelligence Scale*. Next, we will discuss some of the special types of intelligence measures—tests avoiding reading and language, tests for the very young, tests designed to be free from cultural biases. Then we will compare group and individual tests, considering the advantages of each. Finally, the remainder of the chapter will be concerned with evaluation, interpretation, and use of intelligence test results.

### THE REVISED STANFORD-BINET TESTS OF INTELLIGENCE

The individual test that over the years has had the widest use with school-age children is the *Stanford-Binet*, brought out by Lewis M. Terman in 1916. A revised version of the test was published in 1937 by Terman and Merrill, and this has been somewhat further revised in 1960.<sup>21</sup> The current revision, which uses the best items from the two forms of the test brought out in 1937, is known as *Form L-M*. It provides a set of tests for each of twenty levels of ability, starting with tests suitable for the average 2-year-old and going up to four levels suitable for differentiating the abilities of average and superior adults. To illustrate the content of the test, we have picked four levels

at different points on the scale and listed the tests of each level with brief descriptions.

#### TWO-AND-A-HALF-YEAR LEVEL

1. *Identifying Objects by Use.* (Card with 6 small objects attached.)  
"Show me the one that we drink out of." etc.  
Three out of 6 for credit at this level
2. *Identifying Parts of Body.* (Large paper doll)  
"Show me the dolly's hair." etc  
Six out of 6 parts for credit at this level.\*
3. *Naming Objects.* (Five small objects.)  
"What is this?" (Chair, automobile, etc.)  
Five out of 5 for credit.
4. *Picture Vocabulary.* (Eighteen small cards with pictures of common objects.)  
"What's this? What do you call it?"  
Eight out of 18 for credit at this level.\*
5. *Repeating Two Digits.*  
"Listen; say 2." "Now, say 4, 7." etc.  
One out of 3 for credit
6. *Obedying Simple Commands.* (Four common objects on table)  
"Give me the dog," "Put the button in the box."  
Two out of 3 correct for credit.

#### SIX-YEAR LEVEL

1. *Vocabulary.* (Graded list of 45 words.)  
"When I say a word, you tell me what it means. What is an orange?" etc.  
Six words correct to receive credit at this level. Words like tap, gown.\*
2. *Differences.*  
"What is the difference between a bird and a dog?" "Wood and glass?"  
Two out of 3 correct for credit.
3. *Mutilated Pictures.* (Five cards of objects with part missing.)  
"What is gone in this picture?" or "What part is gone?"  
Four out of 5 for credit.
4. *Number Concepts.* (Twelve 1-inch cubes.)  
"Give me 3 blocks. Put them here."  
Four out of 5 different numbers correct.
5. *Opposite Analogies.*  
"A table is made of wood; a window of \_\_\_\_\_"  
Three out of 4 correct for credit.
6. *Maze Tracing.* (Mazes, with start and finish points marked.)  
"The little boy wants to go to school the shortest way without getting off the sidewalk. Show me the shortest way."  
Two right out of 3 for credit.

\* Scored also at one or more other levels.

## TWELVE-YEAR LEVEL

1. *Vocabulary.* (Same as 6-year level.)  
Fifteen words correct for credit at this level. Words like juggler and brunette.
2. *Verbal Absurdities.* (Five statements.)  
"Bill Jones' feet are so big that he has to pull his trousers on over his head. What is foolish about that?"  
Four out of 5 right for credit at this level.
3. *Picture Absurdities.*  
Picture showing person's shadow going wrong way. "What is foolish about that picture?"
4. *Repeating 5 Digits Reversed.*  
"I am going to say some numbers, and I want you to say them backwards."  
One out of 3 correct for credit.
5. *Abstract Words.*  
"What do we mean by pity?"  
Three out of 4 for credit at this level.
6. *Sentence Completion.* (Four sentences with missing words.)  
"Write the missing word in each blank. Put just one word in each."  
Three out of 4 required for credit at this level.

## SUPERIOR ADULT—LEVEL II

1. *Vocabulary.* (Same as 6-year level.)  
Twenty-six words for credit at this level. Words like mosaie, flaunt.
2. *Finding Reasons.* (Two parts.)  
"Give three reasons why a man who commits a serious erime should be punished."  
Both parts right for credit.
3. *Proverbs* (Pearls before swine, etc.)  
"Here is a proverb and you are supposed to tell what it means."  
One out of 2 correct for credit.
4. *Ingenuity.*  
A 5-pint can and a 3-pint can to get exactly 2 pints of water.  
Three out of 3 problems correct for credit.
5. *Essential Differences.*  
"What is the principal difference between work and play?"  
Three out of 3 correct for credit.
6. *Repeating Thought of Passage.*  
Short paragraph on the value of life.  
Four out of 7 essential ideas must be reproduced for credit.

The above examples illustrate the variety of material included in the test. Note that the specific tests vary from one level to another. Many of the tests at the lower age levels are quite concrete, dealing with little objects and pictures. At the upper levels, the tests tend to be more abstract and quite heavily verbal. The various tests include



tasks calling for display of past learnings, perception of relations, judgment, interpretation, sustained attention, immediate memory, and other cognitive processes.

The tasks were selected so as to be of appropriate difficulty for the average child of the age level to which they were assigned. In testing a child, the examiner begins at a level where the child is likely to succeed, but only with some effort. If the child fails these and appears discouraged, the examiner will drop back to an easier level. Otherwise, he will move ahead level by level until he reaches a level at which the child fails all tests. When the upper limit has been established, the examiner will be sure to go back and establish the level at which the child can do all the tasks. Often, a few quite easy tests will be given at the end to build up the child's morale.

The child is credited with the basal age at which he passes all tasks plus a credit for tasks passed at more advanced levels. Each task passed at a given level credits the child with the same number of months of mental age. Thus, where there are 6 tests at each year age level, passing a single test gives a credit of 2 months of mental age. For example, child A

Passed all tasks at 6-year level	= 6 yrs. basal age
Passed 3 of 6 tasks at 7-year level	= 6 mos. credit
Passed 1 of 6 tasks at 8-year level	= 2 mos. credit
Failed all tasks at 9-year level	= 0 credit
<hr/>	
Resulting in a mental age of	6 yrs., 8 mos.

Level of achievement is expressed as a mental age, arrived at as indicated above. The mental age describes the level at which the child is performing. But this takes no account of the child's life age. Performance in relation to a group of children of his own age is expressed as an IQ. The IQ's for this latest revision of the Stanford-Binet are deviation IQ's, i.e., they are essentially standard scores for which the mean is 100 and the standard deviation 16 at each age level. In so far as the normative groups are adequate and comparable from one age to another, an IQ has the same meaning at one age as at any other. Tables for converting MA's to IQ's are provided from age 2-0 (2 years, no months) up to age 16-0. For individuals over 16 years of age the table is entered with a chronological age of 16-0. The way IQ's spread out is shown in Table 6.7 (p. 145). Thus, a child with an IQ of 130 would surpass about 95 per cent (95.5 per cent by Table 6.7) of children of his age; one with an IQ of 90 would surpass about 23 per cent (22.7 per cent by Table 6.7).

## THE WECHSLER INTELLIGENCE SCALES

The second major individual intelligence test is the *Wechsler Adult Intelligence Scale (WAIS)*.<sup>28</sup> This test was originally developed for adults, and the materials and tasks were chosen with an eye to their appropriateness for adults. The pattern of organization of the test differs from that of the *Binet*. Whereas the *Binet*, developed for children, is organized in successive age levels, the *WAIS* is organized by subtests representing types of tasks. The subtests are the following:

*Verbal Subscale*

1. General Information.
2. General Comprehension.
3. Arithmetical Reasoning.
4. Similarities.
5. Digit Span.
6. Vocabulary.

*Performance Subscale*

7. Digit-Symbol Substitution.
8. Picture Completion.
9. Block Design.
10. Picture Arrangement.
11. Object Assembly.

Tasks like those in a number of the subtests will be found among the examples on pp. 222-225.

Each subtest of the *WAIS* yields a separate score, which is then converted into a standard score for that subtest. The subtest standard scores are combined in three different groupings to yield total scores, and from these total scores three different types of IQ's may be read from norm tables. The three IQ's are (1) a verbal IQ from subtests 1 through 6, (2) a performance IQ from subtests 7 through 11, and (3) a total IQ from all the subtests put together. The separate verbal and performance IQ's may have diagnostic significance in the case of certain individuals with verbal, academic, or cultural handicaps. The IQ on the *WAIS* is also a standard score, set to make the mean of the normative sample 100 and the standard deviation 15.

As we have indicated, the original *Wechsler Intelligence Scale* was designed for adults. It was suitable for use with adolescents and with adults of all ages. Subsequently, however, the material has been extended downward to make a test for children.<sup>29</sup> The same general pattern of subtests has been used, though with minor variations. In particular, the nature of the tasks in several of the subtests changes as one goes down to the easiest items. The *Wechsler Intelligence Scale for Children (WISC)* is designed to be usable from age 5 to 15.

The features that distinguish the *WAIS* from the 1937 edition of the *Stanford-Binet* are:

1. Original test items specifically designed for adults.
2. Organization by subtests rather than by age levels.
3. Provision for separate verbal and non-verbal IQ's.

All these features seem like sound adaptations in a test for adults. Most psychometricians would probably agree now in preferring the *WAIS* as a measure for adolescents and adults, though its relation to academic success is perhaps not as clearly established as is the *Binet's*. (As a matter of fact, at these ages a printed group test would usually seem more appropriate for academic prediction.)

The *WISC* cannot be used with children below 5 and is probably not very satisfactory below the age of about 7. For young children the *Binet* would be generally preferred. In the age range from 7 to 15, a decision between the two tests is not an easy one. The *Binet* is reported to be somewhat more difficult and time-consuming to give. The usual *Binet* procedure of carrying the examinee through to the point where he encounters a long series of failures is judged to be a seriously upsetting matter for some emotionally tense children. The separate verbal and performance IQ's of the *WISC* should be quite useful in some cases in understanding children whose verbal development is either very accelerated or retarded. It has diagnostic value for some children with special educational disabilities. However, the *Binet* is probably a somewhat more reliable measure. (No directly comparable data are available.) The test items entering into the *Binet* have had the benefit of trial in earlier forms, with opportunity to revise and select on the basis of that experience. The ultimate basis for choice will be the validity of the inferences that can be made from each in the situations in which they are actually used. Prediction of academic success can apparently be made about equally well from either test. It seems likely that the two tests are about equally useful for children with mental ages of 7 or above.

## NON-LANGUAGE AND PERFORMANCE TESTS

Most of the widely used intelligence tests depend to some degree upon language and include tasks presented in verbal terms. This is natural, since the bulk of our learning and thinking makes use of language. For the usual person and in relation to the usual type of academic learnings, aptitude for learning can be tested more efficiently by tasks that involve language than by those that do not. However, for some groups or situations this is not so. The most obvious example is that of groups who do not speak the language or speak it only slightly. When an individual has limited command of English, results from a verbal test in English are in large measure meaningless. Children who have had little opportunity to attend school may suffer a special handicap on a test that relies upon materials close to school learnings. For groups of this sort, tests have been developed that do

not require language. In some of these, only the test tasks are non-language in character; in others the instructions can be given by pantomime and no language need be used at any point during the testing.

A group test that requires no language in solving the test problems, though the instructions are presented in words, is the *Lorge-Thorndike Intelligence Test, Non-Verbal Series*. Types of tasks that are included are figure analogies, figure classification, and number series. (See examples on p. 223.) A group test that dispenses with language in both instructions and test is the *Pintner Non-Language Test*, in which all instructions may be given by pantomime. The test includes the following types of tasks, which are illustrated in Fig. 9.2.

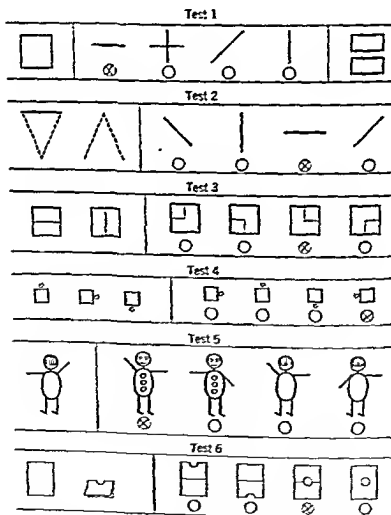


Fig. 9.2. Sample items from Pintner Non-Language Intelligence Test. (Copyright 1941, World Book Co., Yorkers, N. Y. Reproduced by permission.)

1. *Figure dividing*, indicating which line or lines will divide a figure up to give a specified set of parts
2. *Reverse drawings*, indicating the line or lines needed to complete a mirrored drawing.
3. *Pattern synthesis*, indicating the figure that will result from superimposing two figures.
4. *Movement sequence*, selecting the figure that follows the movement sequence established by three figures in the stem of the item.
5. *Manikin*, selecting the manikin that is the same as the one in the stem, except rotated in some way.
6. *Paper folding*, selecting the diagram that shows how a paper folded and cut in a specified way will look when unfolded.

When used with ordinary school groups, a test such as the *Pintner Non-Language* provides an appraisal of intelligence somewhat distinct from that provided by a verbal measure. Thus, this test correlates only about .65 with the *Pintner General Ability Test, Language Series*, a test made up of verbal and arithmetical material. With usual groups, the non-language test may be expected to be somewhat less effective as a predictor of school achievement. The value of the non-language test is for atypical individuals or groups, i.e., the deaf, the foreign born, or the academically retarded.

Individual tests are also available that do not require the use of language. We have already described the *Wechsler Adult Intelligence Scale* and referred to the performance IQ provided by this test. The performance IQ is based upon five subtests that do not require the subject to use language once he has been instructed as to the nature of his task. A performance test that is widely used with children, as a supplement to the *Binet* when a verbal handicap is suspected, or for groups with which the *Binet* would not be appropriate is the *Arthur Point Scale*.<sup>2</sup> We shall describe it in some detail, since it is a good representative of individual performance tests.

The *Arthur Point Scale* consists of two forms, which contain somewhat different tests. Form I has nine subtests, as follows:

1. *Knox Cubes*: The examiner taps four cubes in a specified sequence, and the subject must reproduce the sequence.
2. *Seguin Form Board*: Ten geometric figures are to be placed into the corresponding holes in the board as rapidly as possible.
3. *Two-Figure Form Board*: Cut-up pieces are to be fitted into a square and cross cut out of the board.
4. *Casist Form Board*: Similar to the above, only four figures.
5. *Manikin or Feature Profile* (depending on level): Cut-up figure of man or cut-up face is to be assembled.
6. *Mare and Foal*: Picture has cut-outs that are to be fitted into place.
7. *Healy Picture Completion I*: Picture has square cut-outs, and subject must select the appropriate block to make the most meaningful picture.

8. *Porteus Mazes*: Simple pencil mazes are to be traced without retracing or crossing a line.

9. *Kohs Block Design*: Designs are to be reproduced using colored cubical blocks, like those in sets for children.

*Form II* of the test also uses the *Knox Cube*, *Seguin Form Board*, *Healy Picture Completion*, and *Porteus Mazes*, presenting a different form or a different set of tasks from *Form I*. In place of the other tests, however, it substitutes the *Arthur Stencil Design Test*. In this test, the subject is supplied with a set of colored cards and a set of cut-outs of different designs and colors. The subject is shown a design that can be produced by superimposing certain ones of the cards provided to him. He must select the right cut-outs and background and put them together in the right order to produce the master design.

A point score is allowed the subject for his performance on each subtest of the *Arthur Scale*. The score depends in some subtests upon the speed with which the task was completed, in others upon the correctness of the solution or the number of graded tasks solved. The point credits for the subtests are summed to give a total point score, and this is converted to a mental age equivalent. An IQ is computed by dividing mental age by chronological age. The IQ's appear to have about the same distribution as for the *Revised Binet*.

There have been a number of other attempts to evaluate intellectual ability through performance tasks, ideally ones that would be usable in different countries and different cultures. One of the most widely known is the *Goodenough Draw-a-Man Test*,<sup>12</sup> in which the child is told simply, "Draw a man—the best man you can draw." The performance is scored on completeness and maturity of representation, not on esthetic qualities.

The individual performance test must generally receive the same evaluation as group non-language tests. For an English-speaking person with normal environmental opportunities and without specialized language or reading handicap, it represents a less efficient way of appraising mental development than the more widely used verbal test. However, as a way of checking on whether there is a specialized language handicap it represents a valuable supplemental tool. It makes it possible to check upon individuals who appear retarded on the verbal type of test to see whether the retardation is general or whether it is a localized deficiency in the language area. A performance test such as the *Arthur Point Scale*, which can be given with pantomime instructions, is also useful in testing deaf children, non-English-speaking children, and other types of special groups.

## INFANT AND PRESCHOOL TESTS

The first intelligence tests were made for school-age children. However, it was not long before the theoretical interests of child psychologists and the practical needs of child-care and placement agencies stimulated the attempt to develop procedures for appraising intelligence in preschool children and even in infants. Any appraisal procedures with young children obviously had to be individually administered. Also, they had to be based upon behavior that was spontaneously exhibited by or could be elicited from children of the age being studied. Infant tests, therefore, had to take on a very different character from later appraisals. Arnold Gesell "pioneered in designing tests based on observation of the child's postural, perceptual, manipulative, and social responses. Does he sit up? Stand up? Walk? Will he turn to look at a light? Notice a face? Can he pick up a block? A spoon? A little pellet? By what type of a grasping motion? How does he react to strange adults? To another infant?

Observations of large numbers of infants showed a typical developmental sequence in the different aspects of the child's development. Performance B followed A, and was followed by C. Norms have been established representing the average age at which a particular behavior manifests itself. The child may be assigned a developmental age, based upon the behavior he shows. Retests after a short interval show the child to be fairly consistent in his level of performance. If he is advanced at one testing, he will tend to be advanced at the other. The developmental schedules provide a moderately reliable picture of the individual at that point in time.

What significance does acceleration or retardation in development during the first year or so of life have for predicting later intelligence? The answer is well presented in Table 9.1, which shows the correlation of infant tests given at the ages of 1 to 12 months with intelligence tests at various later ages. The tests during the first 15 months were those of the *California First-Year Mental Scale*, those from 18 months to 5 years were the *California Pre-school Scale*, and those from 6 years on were the *Stanford-Binet*.

The picture seems quite clear. The infant tests give a fairly good prediction of developmental status a few months later, but their value as predictors drops rapidly as the interval increases. The infant tests provide essentially no prediction of intellectual status at school age. Whatever factors produce differences in rate of development during the first year or so of life are entirely distinct from those that deter-

Table 9.1. Correlation of Intelligence Tests During First Year of Life with Later Measures \*

(Correlations based on pooling of successive tests)

Age at Later Test	Age at Initial Test			
	1, 2, 3 mos.	4, 5, 6 mos.	7, 8, 9 mos.	10, 11, 12 mos.
4, 5, 6 mos.	.57			
7, 8, 9 mos.	.42	.72		
10, 11, 12 mos.	.28	.52	.81	
13, 14, 15 mos.	.10	.50	.67	.81
18, 21, 24 mos.	-.04	.23	.39	.60
27, 30, 36 mos.	-.09	.10	.22	.45
42, 48, 54 mos.	-.21	-.16	.02	.27
5, 6, 7 yrs.	-.13	-.07	.02	.20
8, 9, 10 yrs.	-.03	-.06	.07	.19
11, 12, 13 yrs.	.02	-.08	.16	.30
14, 15, 16 yrs.	-.01	-.04	.01	.23
17, 18 yrs.	.05	-.01	.20	.41

\* Tests used were: 1-15 months, *California First-Year Mental Scale*; 18 months-5 years, *California Pre-school Scale*; 6 years and older, *Stanford-Binet*. From Bayley.<sup>3</sup>

mine intellectual level at school age. It seems, then, that little practical significance can be attached to results from infant developmental schedules. They describe an aspect of the child which is temporary only, not lasting.

There have been a number of different tests prepared primarily for use with preschool children, i.e., the age range from about 18 months to 5 years. As a matter of fact, as we have seen, the *Stanford-Binet Intelligence Scale* has tests going down to the 2-year level and may be considered a preschool test. It would compare very favorably with the other tests available for this age level, though it is somewhat more verbal than many of the others. A good many of the preschool tests have tended to get away from the verbal material that appears so heavily in group tests for older children and also in the *Stanford-Binet*.

One test for preschool children that has received wide use is the *Merrill-Palmer Scale*.<sup>24</sup> This is most suitable for children from 2 to 4, though it can be used with children slightly older and slightly younger. The test is made up of 38 little subtests, of which only 4 call for verbal response by the child. A number of the tasks call for



gross motor coordination (standing on one foot) or finer eye-hand coordination (building block tower, cutting with scissors). Form and object perception and motor control combine in a number of form-boards in which cut-outs must be fitted into the appropriate hole. The tasks make use of a variety of materials interesting to the child, blocks, pictures, scissors, balls, etc., so that cooperation can usually be obtained, a real problem with children at these ages.

The *Merrill-Palmer Scale* has fairly satisfactory reliability, especially above about 30 months. Correlations with retests 6 months later have been reported<sup>9</sup> as follows for different age groups:

24 months	.63
30 months	.76
36 months	.78
42 months	.80

The correlation with school-age *Binet* is about .40 for a *Merrill-Palmer* test at age 2; about .45 to .50 for one at age 4.

The *Minnesota Preschool Scale*<sup>10</sup> is another example of a test designed for preschool groups. The 26 tests in this scale tend to be more like those of the *Binet*. Six tests taken at random from one form of the Scale are described briefly. They are

*Test 2: Pointing Out Objects in Pictures.* Card with man, chair, apple, house, and flower on it. Child is asked to point to each in turn.

*Test 5: Imitative Drawing.* Experimenter makes vertical stroke; then a cross. Child is asked to imitate each in turn.

*Test 8: Imitation.* A set of 4 cubes, on which experimenter taps in specified sequence. Child instructed to imitate the sequence of taps.

*Test 14: Colors.* Cards colored red, blue, pink, white, and brown. Child is asked to name the color.

*Test 20: Paper Folding.* Examiner folds paper with three consecutive folds. Child is asked to copy exactly.

*Test 24: Giving Word Opposites.* Child is asked to give words meaning opposite of cold, bad, thick, dry, dark, and sick.

Test materials are quite simple. Copying, imitating, and responding to simple verbal relations enter into a number of the tests.

This test appears to be somewhat more reliable than the *Merrill-Palmer*. Correlation between two forms of the test given within a few days of each other was found to be .89. Below 3 years, this test did not correlate very well with later *Binets*, but the *Minnesota* given between 3 and 4 gave a correlation with *Binets* at school age of about .60. However, IQ's on the *Minnesota Preschool Scale* have quite a different spread from those for the *Binet* so a preschool IQ on this test is not readily equated to later *Binet* performance. (See reference 14.)

## CULTURE-FREE AND CULTURE-FAIR TESTS

Many workers in the field of aptitude testing have been distressed by the fact that test performance depends upon the experiences the person has had. Every test maker has recognized this to a degree and has tried to base test items upon experiences that would be common to the group for whom the test was planned. But some have perhaps taken too narrow a view of the group for whom the experiences should be common. Certainly the test that incorporates pictures of the usual American house, automobile, or football is not suitable for an Australian Bushman who has seen none of these objects. The typical American test assumes the common core of an American culture. Some critics have gone further and asserted that the typical test is based upon an urban middle-class American culture. Both in its highly verbal content and its emphasis upon speed, competition, and doing one's best, it is said to be centered in the middle-class culture and values.

Several attempts have been made to develop tests that are "culture free," or if not that at least "culture fair." These are closely related to the non-verbal and performance tests described in the previous section, because a culture-free test is almost necessarily non-verbal. It must not only be non-verbal but must also be free of the content of any particular culture.

One attempt to develop such a test is the *Cattell Culture Free Intelligence Test*. The *Cattell Test* is based on the premise that general intelligence is a matter of seeing relationships in the things with which we have to deal, that the ability to see relationships can be tested with simple diagrammatic or pictorial material, and that for a test to be usable in different cultures the pictures should be of forms or objects which are fairly universal, i.e., not peculiar to any cultural group. Items illustrating the different types of tasks are shown in Fig. 9.3. The evidence that the test is in fact useful for widely different cultures is largely lacking, but the tasks constitute one further interesting non-verbal group test that may prove usable, particularly in research studies.

One test that was developed in Great Britain and has been used in many countries is the *Progressive Matrices Test*.<sup>20</sup> The type of item is similar to the last two samples in Figure 9.3. Two types of progression or relationship are established, one in the horizontal and one in the vertical direction. The examinee is required to pick the choice that correctly fills the missing entry in the lower right-hand corner of the matrix.

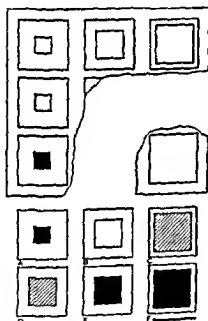
## PART I-CLASSIFICATIONS



## PART III-SERIES



## PART V-MATRICES II



## PART VI-MATRICES III

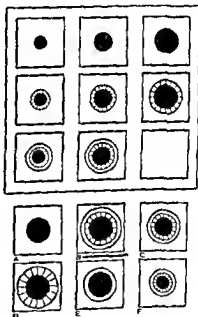


Fig. 9.3. Sample items from Cattell Culture-Free Intelligence Test. (Copyright 1944, Institute for Personality and Ability Testing, 1602 Coronado Drive, Champaign, Ill. Reproduced by permission.)

An attempt to develop a test that imposes no penalty on different classes in American society is found in the *Davis-Eells Games*. (Presumably the child is supposed to be naive and not realize that this is a test.) This test series involves no written language but does require quite long oral directions. Types of items include:

1. *Best ways*, in which three pictures are shown in the test booklet, and the examinee is orally instructed to mark the one that is the best way to carry a pile of packages, get over a fence, etc.

2. *Analogies*, in which the analogies are presented in pictures and are of the type, "Glove is to hand as sock is to: arm, leg, foot."

3. *Probabilities*, in which a picture is shown and the examinee must select the one of three orally presented choices that indicates what probably led up to or is represented in the picture.

4. "*Money*," a task based on complex directions for following certain rules for combining coins to make specified sums.

This test was designed to avoid the cultural biases, particularly socio-economic biases within the American culture, thought to characterize previously existing tests. However, studies of the test in recent years have failed to confirm that it does so. IQ's from the *Davis-Eells Games* are found to have about as high a correlation with indices of socio-economic status as those for any other test. We must conclude that either there basically is a relationship between mental ability and socio-economic status, or that the *Davis-Eells Games* has failed to eliminate the bias which its authors believed to characterize other tests. Since this type of test is laborious to give and relatively unreliable, it has little to recommend it on other grounds. We must conclude that it does not appear very useful as a measurement tool at the present time.

## GROUP VERSUS INDIVIDUAL TESTS AS MEASURES OF INTELLIGENCE

We have seen that intelligence tests fall into two main patterns, group tests and individual tests. The types of tasks presented to the examinee are a good deal alike in both patterns. However, the two procedures have certain significant differences. These may be summarized as follows:

### *Group Tests*

Problems presented in printed booklet. Read by examinee. Personal contact with examiner a minimum.

Tasks presented and test timed as a unit, or separate time limits for each subtest.

Individual usually responds by selecting one of a limited set of response options printed in the test booklet.

### *Individual Tests*

Problems presented orally by examiner in face-to-face situation.

Problems presented one at a time, usually without indication of time limits.

Individual usually responds freely, giving whatever response seems appropriate to him.

These differences in procedure have several important implications for the conduct of testing and for the results that may be obtained from such testing. In the first place, when test tasks are presented orally to the subject and he does not have to read them for himself, his performance is much less dependent upon his reading skills. The child who has lagged behind in acquiring these skills is not penalized for this specific failure. The effect of reading disability upon intelligence test performance is shown clearly in a study<sup>4</sup> comparing individual *Stanford-Binet* scores and group-test scores of retarded, normal, and accelerated readers in the sixth grade. For those children whose reading was a year or more accelerated (in relation to *Stanford-Binet* mental age), group-test IQ averaged 15 points higher than the individual *Stanford-Binet* IQ. Where reading was within + or - 1 year of *Stanford-Binet* mental age, the group test IQ was 2 points higher. Where reading was retarded a year or more, group test IQ fell 8 points below the *Stanford-Binet* IQ. Thus, in this study the accelerated reader received a 15 point bonus, the retarded reader an 8 point penalty in IQ on the printed group test as compared with the individual test orally administered.

The results reported above are probably somewhat extreme, because the particular group test was very verbal in nature and because the study was carried out with elementary-school children, for whom the actual operation of reading still represents something of a task. One may anticipate that less difference would be found for high-school or college students. Furthermore, some current group tests are either partly or wholly non-language in their content and would be relatively independent of reading skills. However, this study points out very clearly the caution with which a group test IQ must be interpreted for a person who departs markedly from the average in his reading skills. A low group test IQ for a poor reader cannot be taken at face value. It should always be checked with a test that does not involve reading.

The presentation of problems one at a time by an examiner is also a factor of some significance in determining what the test is likely to yield. Especially with younger children, maintaining continuity of attention and effort on a group test may be a problem, and variations in this respect are certainly a significant factor in test score. When each problem is separately presented by the examiner, this serves to re-establish the child's orientation to the task and to maintain his effort. What is equally important, the examiner is in a position to observe lapses of interest and effort and to take some account of them in interpreting the results.

The individual intelligence test is essentially a well-standardized interview situation. The tasks to be presented to the examinee are

specifically formulated, and detailed standards are provided for evaluating his responses. However, at the same time, the face-to-face relationship of an interview prevails. This offers the alert examiner a wealth of opportunities for observing the examinee and noting poor motivation, distractability, signs of anxiety and upset, and other cues that will help in interpreting the actual test performance. At the same time, the demands upon the examiner are considerably heavier. If valid testing is to result, the tasks must be presented in a standard way, interest and cooperative effort must be maintained, and a uniform standard must be applied in evaluating responses.

The free-response item in the individual test fits into the interview setting of the individual test and reinforces both its strengths and its limitations. Potentially, the free response of the examinee can tell us more about him than the mere record of which option he has chosen from a set of five. There is more of the quality of his own behavior available to us. We can see just how he goes about defining a word, whether by class and differentia (i.e., an orange is a round, orange-colored, citrus fruit) or by use (an orange is to eat). We can note the speed and sureness of his attack on a problem task. But we must also depend on the examiner to interpret and evaluate the responses, and at this point subjectivity is likely to creep into the examining. Careful attention must be paid to the standard samples provided in the test manual, and experience under supervision is indicated before an examiner can expect to give and score an individual intelligence test in a way that will yield results comparable to those of other examiners.

In general, the limitations of group tests are most acute and the advantages of individual tests most pronounced with young children. Printed group tests cannot be used successfully with children below school age. They cannot read and have difficulty in manipulating a pencil, following instructions, or maintaining sustained attention for the period that is required for taking a test. These same factors continue to present fairly serious problems for testing in the primary grades. However, the factor of cost makes individual testing impractical for most large-scale users of tests, so that with older individuals the overwhelming majority of the intelligence tests used are paper-and-pencil group tests.

## RELIABILITY AND STABILITY OF MEASURES OF INTELLIGENCE

We have already presented some evidence on the reliability of measures of intelligence in our discussion of infant and preschool tests. The reliability of those early measures is found to be quite modest.

For tests at school age, reliabilities are more promising. Considering the group tests first, we find that when correlations between two forms of the same test are reported for an age group or a grade group they usually fall between .80 and .90. A few are higher. Unfortunately, the authors of some tests report only odd-even reliabilities, and it is difficult to estimate how much these are inflated. (See discussion on pp. 180-181.) Comparisons of different tests are made difficult by variations in the procedure used for estimating reliability and in the type of group for which results are reported.

The correlations reported by the authors<sup>23</sup> between Form L and Form M of the *Stanford-Binet Intelligence Scale* ranged from .85 to .95 for different age groups. For ages from 2 to 6, the median value was .88, whereas for ages above 6 the median was .93.

Since Form L-M was prepared by selecting the better items from both Form L and Form M, one may anticipate that the reliability of the new form is at least as high as these values. The reliability reported in the manual for the *Wechsler Adult Intelligence Scale* is .96 for the verbal IQ, .93 for the performance IQ, and .97 for the full scale IQ. These are split-half reliabilities, and consequently should be discounted somewhat in relation to the reliability of the *Binet*. Split-half reliabilities for the *Wechsler Intelligence Scale for Children* are reported to be .92 at age 7½, .95 at age 10½, and .94 at age 13½.

Though the variations in procedure for estimating reliability and in type of group tested make it difficult to arrive at an unequivocal answer, it does seem that the individual intelligence tests yield a somewhat more reliable measure than do the commonly used group tests. This is probably in part a reflection of the somewhat longer actual testing time, in part a result of more uniform motivation and effort when working under the eye of the examiner.

The reliabilities of intelligence tests are reasonably satisfactory, and they are among the most dependable psychological measuring instruments. However, the chance errors in an IQ are still enough to require that we be quite tentative in our interpretation. Thus, Table 9.2 shows the spread of IQ's that could have been expected on Form M of the *Binet* if that form had been given to a group of pupils all of whom had received exactly the same IQ on Form L. Note that the IQ's spread over a range of more than 25 points, and that less than a third of the cases fall in the center 5-point interval. And it must be remembered that these figures are for the *Stanford-Binet*, one of our most reliable tests. Thus, an IQ of 100 must not be thought of as meaning "exactly 100," but rather "probably between 95 and 105, very probably between 90 and 110, almost certainly between 85 and 115."

Table 9.2. Distribution of Stanford-Binet Form M IQ's for Cases with Identical Form L IQ's

IQ	f
113+	3
108-112	9
103-107	23
98-102	30
93-97	23
88-92	9
87 and below	3

## STABILITY OVER A PERIOD OF YEARS

In addition to knowing the precision with which an intelligence test appraises an individual's abilities at a particular time, we would like to know how consistently the individual maintains his position in his group from one year to the next or over a considerable span of years. How confidently can we predict what scholastic aptitude an individual

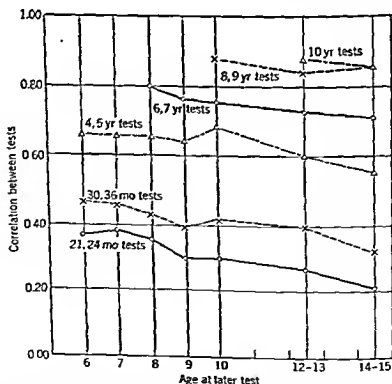


Fig. 9.4. Effect of age at initial testing and test-retest interval on prediction of later Stanford-Binet IQ from earlier test. (Adapted from Honzik, McFarlane, and Allen.<sup>16</sup>)



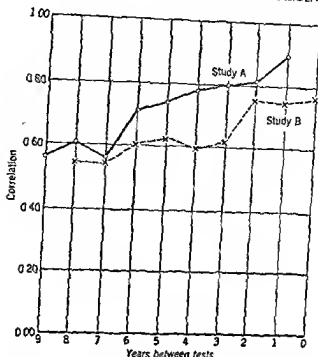


Fig. 9.5. Effect of test-retest interval on prediction of group test intelligence at end of high school from earlier group tests. [Study A adopted from J. E. Anderson,<sup>1</sup> Study B adopted from R. L. Thorndike,<sup>20</sup>]

will show when he is of college age from his performance on a test at age 2? Age 6? Age 10? Evidence on this point is presented in Figs. 9.4 and 9.5.

Figure 9.4 shows the findings from one extensive study using individual tests. The final test is the *Stanford-Binet* in every case. The initial test is the *California Pre-school Scale* up through 5 years and the *Stanford-Binet* after that age. Note that for the early tests the prediction is rather poor and drops as the interval is increased. A test at age 2 correlates only .37 with one at age 6 and .21 with one at age 14 or 15. As we go up the age range, however, the correlations are higher and the drop is less. A test given at age 8 or 9 correlates .88 with one at age 10 and still correlates .86 with one at age 14 or 15. For normal children in a typical environment, a *Stanford-Binet* at age 8 or 9 appears to provide almost as accurate a forecast of ability near the end of high school as would the same test given several years later.

Two sets of data on stability of group-test performance over time are presented in Fig. 9.5. The two follow the same general pattern, though they differ a good deal in detail. As we go back further in

time, the correlation coefficients tend to drop more or less steadily. The earlier tests at around grade 3 or 4 correlate perhaps .50 to .60 with the final test, but for a test in grade 9 or 10 the correlation is .70 to .80. In these studies of group tests, the tests that were used differed at the different ages. For this reason, it is not clear how much the lower correlation over the longer intervals is due to growth changes in the subjects over a span of years and how much it is due to changes in the material included in the tests. From the practical point of view, Fig. 9.5 suggests that a group intelligence test needs to be supplemented by new testing every 3 or 4 years if pupil records are to provide an accurate indication of current ability level.

### THE PRACTICAL IMPORTANCE OF INDIVIDUAL DIFFERENCES IN MEASURES OF INTELLIGENCE

To what extent are the individual differences that are brought out by tests of intelligence of importance in the practical affairs of life? Do they enable us to predict to a useful degree how an individual will perform in school, on a job, or in other life adjustments?

#### INTELLIGENCE AND SCHOOL SUCCESS

First, let us consider academic success. From the many hundreds of investigations of intelligence test scores in relation to academic success, a number of conclusions can safely be drawn. These may be summarized as follows:

1. *The Correlation of Intelligence Test Score with School Marks Is Substantial.* Viewing all the hundreds of correlation coefficients that have been reported, a figure of .50 to .60 might be taken as fairly representative. Though this constitutes a very definite relationship, it is only necessary to turn back to Fig. 5.7 and the discussion of correlation on p. 119 to realize that there are still many marked discrepancies between intelligence test score and what a particular youngster does in school.

2. *Higher Correlations Have Been Found in Elementary Schools Than in High Schools and in High Schools Than in Colleges.* Past studies have indicated a drop in correlation from perhaps .70 in elementary school to .60 in high school and .50 in college. The drop in correlation is probably to be explained by the decreased range of intellectual ability in the college groups. A relatively small percentage from the lower half of a school population go on to college, and specific colleges draw from an even more restricted ability range. Though

more and more young people are going to college, the clientele of specific colleges continues to be fairly homogeneous in ability.

3. *Previous School Achievement Has Given Correlations with Later School Success as High as or Higher Than Intelligence Test Score.* In predicting college marks, for example, high-school record has usually shown correlations at least as high as those resulting from a scholastic aptitude test at entrance.

4. *Intelligence Test and Achievement Combined Give Still Better Prediction.* By pooling information on previous school achievement and intelligence test score, the correlation with later school achievement can be raised above that yielded by either factor alone. The two types of information supplement one another.

5. *Intelligence Tests Correlate Higher with Standardized Measures of Achievement Than with School Marks.* Correlations between an intelligence test and total score on an achievement battery in the .70's or even .80's are not unusual. Thus, for one large eleventh-grade group the correlation between the *California Test of Mental Maturity* and total achievement on the *Progressive Achievement Test* was found to be .71,<sup>4</sup> whereas for a group from grades 4, 5, and 6 it was .84. Another report<sup>7</sup> gives correlations of .84 and .78 for grades 5 and 7 for the correlation between the *Pintner General Ability Test* and the *Metropolitan Achievement Test*.

6. *The Degree to Which Intelligence Tests Are Related to Academic Success Depends Upon the Subject Matter.* As one would expect, the more academic subjects, which depend more completely upon the same kinds of verbal and numerical symbols as those that bulk so large in intelligence tests, show the higher correlations. Thus, one summary of studies in secondary and higher education<sup>21</sup> reports an average correlation of .46 with natural science grades and .38 with English grades and foreign language grades but only .28 with shop work and .22 with grades in domestic science.

The fact that intelligence tests correlate with academic achievement and school progress is unquestioned. From the very way in which the tests were assembled it could hardly be otherwise. How these facts should be capitalized upon in educational planning and individual guidance is a more troublesome matter. We will return to it later in the chapter.

#### INTELLIGENCE IN RELATION TO OCCUPATIONAL LEVEL

We turn our attention now to out-of-school accomplishments and consider how intelligence test scores relate to achievement in the world of work. There are two types of questions that we may raise: (1)

How do workers in different kinds of jobs compare in measured intelligence? (2) Within a given kind of job, to what extent is intelligence related to job success?

In relation to the first question, we have a good deal of evidence stemming from the testing of recruits carried out during World Wars I and II. Data for a selection of representative jobs are shown in Table 9.3. This table shows the 10th, 25th, 50th, 75th, and 90th per-

Table 9.3. AGCT Standard Scores of Occupational Groups in World War II

Occupational Groups	Percentile				
	10	25	50	75	90
Accountant	114	121	129	136	143
Teacher	110	117	124	132	140
Lawyer	112	118	124	132	141
Bookkeeper, general	108	114	122	129	138
Chief clerk	107	114	122	131	141
Draftsman	99	109	120	127	137
Postal clerk	100	109	119	126	136
Clerk, general	97	108	117	125	133
Radio repairman	97	108	117	125	136
Salesman	94	107	115	125	133
Store manager	91	104	115	124	133
Tool maker	92	101	112	123	129
Stock clerk	85	99	110	120	127
Machinist	86	99	110	120	127
Policeman	86	96	109	118	128
Electrician	83	96	109	118	124
Meat cutter	80	94	108	117	126
Sheet metal worker	82	95	107	117	126
Machine operator	77	89	103	114	123
Automobile mechanic	75	89	102	114	122
Carpenter, general	73	86	101	113	123
Baker	69	83	99	113	123
Truck driver, heavy	71	83	98	111	120
Cook	67	79	96	111	120
Laborer	65	76	93	108	119
Barber	66	79	93	109	120
Miner	67	75	87	103	119
Farm worker	61	70	86	103	115
Lumberjack	60	70	85	100	116

centiles on *Army General Classification Test* standard score (based on standardization with an average value of 100 and a standard deviation of 20). A marked gradient is noticed from such occupations as accountant, teacher, and lawyer to such occupations as barber, miner, and lumberjack. The gradient follows fairly closely the educational requirements or average educational background for each occupation. In general, one may say that occupations select out individuals jointly on the basis of educational level and of intelligence. Whether intelligence enters as a significant factor excepting as it determines educational level is more difficult to determine. In any event, the net result is appreciable difference between different occupational groups in performance on intelligence tests.

While noticing the differences between groups, one must not forget the substantial range of score within each group. Individuals differing widely in abstract intelligence function together in the same occupation. Thus, the upper 10 per cent of meat cutters did as well on the *AGCT* as the average lawyer. The bottom 10 per cent of lawyers showed no more intellectual ability than the upper 10 per cent of miners. In spite of group differences in average score, there are still wide *individual differences within groups*.

#### INTELLIGENCE AND JOB SUCCESS

What can we say about the relationship of intelligence test score to success within particular jobs? A summary of the findings reported in a number of different studies is presented in Table 9.4. With the

Table 9.4. Relationship of Intelligence Test Score to Measures of Job Success

Type of Job	Median Correlation with Job Success	Per Cent Significantly Positive Correlations *	Number of Coefficients
Clerical workers	.35	70	84
Supervisors	.40	78	9
Salesmen	.33	100	4
Sales clerks	-.09	6	18
Protective services	.25	33	6
Skilled workers	.55	100	6
Semiskilled workers	.20	47	45
Unskilled workers	.08	31	13

Adapted from E. E. Ghiselli and C. W. Brown.<sup>12</sup>

\* Significant at 5 per cent level.

exception of sales clerks, the median correlation is positive in each case. But for unskilled and semiskilled workers the correlations are quite small. They are higher for clerical workers, supervisors, and skilled workers, though only in the case of the skilled workers are they as high as the typical correlations with school success. In part this may be due to limitations in the *criterion* of job success. Whether success is measured by supervisors' ratings, as is usually the case, or by some index of production on the job, the indicator is likely to be unreliable and biased by a number of considerations that have nothing to do with the real efficiency of the worker. In so far as this is true, no test given to the individual can be expected to predict the criterion.

All in all, we may conclude that (1) intelligence is related to occupational group membership and (2) though the relationship of intelligence test score to job success is usually positive, it is likely to be quite low. Prediction of out-of-school achievement appears a good deal less accurate than prediction of school achievement.

### INTERPRETATION OF GROUP DIFFERENCES IN MEASURED INTELLIGENCE

As soon as the first intelligence tests were developed, investigators started administering them to different kinds of groups and studying group differences in performance on the tests. They compared the sexes, different age groups, groups of different racial or national origin, urban and rural groups, groups from different parts of the country, groups from different socio-economic levels, and so forth. The findings from these studies were fairly consistent in showing appreciable group differences. Lower score on intelligence tests was associated with lower socio-economic status, living in a rural area, living in the Southern or Southwestern United States, being an Indian or Negro, being in an immigrant family from the south of Europe, or being over 40 years old. However, the interpretation of these findings has been a source of a good deal of confusion and conflict.

The first naive tendency was to interpret group differences in intelligence test performance as an indication of innate hereditary differences between the groups in question. For example, the lower test performance of the children of laboring class parents was interpreted as indicating basic genetic differences between that group and the white-collar group. Now, such basic genetic differences have not been *disproved*, but many lines of evidence have made psychologists much more cautious in interpreting group differences in intelligence test performance. Many studies have pointed out the role of life experience

in influencing test scores and have made us realize how dangerous it is to make any comparison of groups whose experiences differ radically. We shall consider some of the relevant evidence.

The testing in the United States in World War I and in World War II has made possible a comparison of the level of performance of the military recruit population in 1918 with that in 1940 to 1945. Using a somewhat revised edition of the 1918 *Army Alpha Test* with a sample of World War II recruits, it was possible to estimate the *Army General Classification Test* equivalents of different scores on *Army Alpha* and thus to compare the performance of the two recruit populations. It was found<sup>21</sup> that the average World War II recruit surpassed 83 per cent of the World War I group.

A similar comparative study, on a smaller scale, was made of children in certain mountain counties of eastern Tennessee.<sup>20</sup> When 1940 performance was compared with that in 1930, it was found that the average IQ for children in these counties had risen from 82.4 to 92.2, a gain of 9.8 points. This gain paralleled a very considerable increase in accessibility and cultural opportunities in the counties in question.

Comparisons of national groups in their own countries have failed to substantiate differences found between immigrant groups in the U. S.<sup>19</sup> Studies of Negro children in New York City have shown a tendency for the IQ's to be higher for those children who had spent a longer time in New York.<sup>18</sup> Studies of foster children have found a level of intelligence for these youngsters above what would have been predicted from the intelligence or social level of their biological parents.<sup>22</sup>

All these findings point to the fact that intelligence test score depends upon experience. Where groups differ widely in experience, differences in test score may be expected to result. Thus, in the United States between 1918 and 1940 the median schooling of 18-year-olds increased from about 8½ years to about 10½ years. In addition, radio sets appeared in over 80 per cent of the homes of the country. Good roads pushed out into the rural areas, so that it was relatively easy to get to town. These are only some of the social and cultural changes. These changes had their impact upon test performance. A more educated population, exposed to more experiences and perhaps especially to more extensive and varied use of language, did better on the tests.

The present discussion does not negate the significance of intelligence test differences in *individuals*. These differences are large even for individuals who have had closely similar environmental opportunities. Environment and experience are not the whole story or per-

haps even a major part of the story. However, the discussion should make us slow to accept group differences uncritically on their face value. It should also make us realize that in interpreting the performance of an individual, some allowance must be made for the environmental opportunity he has had. An IQ of 90 has a rather different meaning for a Negro child who spent his early years in a share-cropper's cabin in the rural South from what it has for the son of the local banker.

## USING INTELLIGENCE TEST RESULTS IN SCHOOLS

There are, in general, three types of settings in which standardized tests are used in schools, and intelligence tests should be considered in relation to each of these. Standardized tests may enter into administrative policy as a basis for administrative decisions on such matters as class grouping, promotion, eligibility for certain classes and curricula, and the like. Standardized tests may be used by the classroom teacher as aids to understanding the individual pupils with whom he must deal and in making adaptations and adjustments to their individual needs. Tests may be used by the guidance staff of the school in planning the most effective use of special resources for diagnostic and remedial teaching, in helping the pupil and his family arrive at sound and realistic educational and vocational plans, and in helping understand personal adjustment crises when they arise. We may consider intelligence tests in each of these contexts.

### INTELLIGENCE TESTS AND THE SCHOOL ADMINISTRATION

Intelligence tests are likely to enter into the actions of the school administration either (1) through a policy of using test results as one basis for forming the group for a classroom or (2) through regulations specifying score levels that permit or require some special action, e.g., assignment to a slow-learning class, eligibility to take algebra, eligibility for a special school, etc. What is an appropriate attitude toward administrative actions of these sorts?

*Grouping by Intellectual Ability.* The policy of forming class groups at least in part on the basis of the intellectual level of the pupils remains a common one. In 1947 to 1948 more than half of city school systems reporting<sup>19</sup> used ability grouping in some form in one or more schools. However, the procedure remains a controversial one. In part this is due to the varied and somewhat contradictory results obtained in studies of the effects of ability grouping.<sup>5</sup> In part it is due to the variety of specific practices subsumed under the same label of



"ability grouping" or "homogeneous grouping." In part it is based upon the different initial biases of those discussing the problem.

It is probably impossible to make any single general evaluation of ability grouping that would apply to all instances of the practice. It can be pointed out that grouping together pupils of like mental ages is only a *first step* to permit adapting class program and procedures to the abilities of the pupils in the class. What is most important is the adaptations that are actually made in materials and procedures after the grouping has been carried out—and also what attitudes exist or can be developed in the community toward the grouping and the adjustments that accompany it. It should also be noted that groups formed on the basis of intelligence-test scores will still be quite heterogeneous with respect to academic skills. The correlations of intelligence and achievement, and of different aspects of achievement are low enough so that forming groups on any one measure will still leave quite a range of performance on any of the others. In a departmentalized program, as in high school, effective grouping in separate subject areas can be based on a combination of an intelligence test and a measure of achievement in the subject area. Though a general evaluation of achievement can be combined with intelligence test score for elementary school pupils, it is not possible to get a group homogeneous for all subject areas.

Many of both the gains and hazards of ability grouping have been claimed to lie in relatively intangible areas of interest, attitude, and adjustment. Evaluations in these areas have generally been quite inadequate. Thus, it is still largely a matter of opinion whether the bright child develops better work habits and leadership traits or feelings of snobbishness and superiority from being in a special class group.

Ability grouping for the bulk of pupils is one issue, and special classes for the relatively extreme deviate is a somewhat different one.

How about the highest and lowest 2 or 3 or 5 per cent in intelligence? Here we must recognize that special administrative provisions are possible only in a community of some size. Unless there are perhaps 500 children per grade in the school system, there will not be enough extreme deviates to fill a class group. The problem of the extreme deviate becomes most acute in the case of the low deviate, because of the obvious problems that the slow learners have in adapting to the activities and tempo of a regular classroom. Special class groups have not been a universal panacea, but they do permit adaptation of the type of class activities and the rate of progress to the interests and abilities of the slower learners.

The very bright child is usually a less conspicuous problem in the

regular class. He gets the regular work done. His boredom is less apparent. Furthermore, the alert teacher can often provide supplementary activities which will keep him profitably occupied. However, there is evidence<sup>17</sup> that children of high ability who are placed in special groups can master the regular school curriculum more rapidly than they would in regular classes, or engage in a wide range of enrichment activities without falling behind children in regular schools. Furthermore, there is no real evidence that membership in special class groups results in undesirable personality attributes in these children. In view of the importance of individuals of high ability for our society and in view of the long period of training that most of them must undergo to take a role in the professional groups of our society, special provisions to accelerate or enrich their early training would seem to be a sound social provision where such provisions are administratively feasible.

*Intelligence Test Score as an Administrative Prerequisite.* Intelligence test results enter into administrative actions when a certain level of intelligence is specified as a prerequisite for some action in relation to a pupil. Generally speaking, the relationship of intelligence test score to educational progress or success is low enough and the variety of factors involved is great enough so that rigid administrative standards on intelligence seem rather questionable. Intelligence is often a factor that should receive consideration, together with other factors, in arriving at a decision with respect to any individual. But room for flexibility of action is needed, in the light of all relevant factors. An administration should formulate general policy with respect to the use of intelligence tests for admitting pupils to special groups, but the policy should be one which permits actions on individual cases to be taken in the light of a variety of relevant factors.

#### INTELLIGENCE TESTS AND THE CLASSROOM TEACHER

The classroom teacher will want to use intelligence test results as an aid to understanding each pupil in the class and to providing the school experiences that will be most helpful to that pupil. The child's level as measured by an intelligence test provides probably the best single clue available to the teacher as to the child's potentialities for learning the abstract symbolic aspects of the school curriculum. The test results provide a guide as to what can reasonably be expected of each pupil: whether the pupil should be expected to move along as rapidly as the rest of the class, whether the pupil's achievement is falling enough behind expectation to suggest the need for special diagnostic or remedial procedures, or whether the pupil's abilities are

enough ahead of those of the bulk of the class so that the teacher should try to provide special activities and opportunities for enriching the regular program.

There are certain cautions that need to be observed when the classroom teacher makes use of intelligence test scores for his pupils. An enumeration of the pitfalls may help the reader to avoid them.

1. The general intelligence test, especially the group test, is a measure of ability to work with symbols, abstract ideas, and their relationships. This is one quite limited type of ability. The test does not encompass ability to work with things or people, or perhaps the ability to solve many types of concrete and practical problems. The child who is low on an intelligence test will probably have trouble with the academic aspects of the conventional school curriculum. However, he may have a good level of skill or ability in the many non-abstract aspects of living—mechanical, social, artistic, musical. The teacher should seek these strengths, capitalize upon them, and build upon them. *Above all, the teacher must recognize that intelligence test score is not a measure of personal worth and must avoid rejecting the child whose aptitude for academic pursuits is low.*

2. The verbal group intelligence test that is ordinarily used for school-wide testing is sufficiently dependent upon reading and arithmetical skills that a low test score must be interpreted cautiously for a poor reader or low achiever in arithmetical skills. If possible, individuals of this sort should be tested also with an individual test or a non-verbal group test to determine whether the low performance is due to limited ability, or whether it is a reflection of limited reading and number skills.

3. Intelligence test results for a child whose social and cultural background differs radically from that of the rest of the group should be interpreted with caution. The possibility of some degree of environmental deprivation should be borne in mind.

4. If it is known or suspected that a child was emotionally disturbed at the time of testing, results should be considered quite tentative. Motivation and effort are needed for sound test results.

5. The standard error of measurement should always be very real to the test interpreter. An IQ of 90 should always signify to the teacher "IQ somewhere between 80 and 100."

#### INTELLIGENCE TESTS AND THE GUIDANCE STAFF

Intelligence tests have their most obvious function in the educational program as sources of information important to persons responsible

for counseling and helping the child with problems of personal and social adjustment, making provisions for special educational activities for him, helping him to decide on appropriate educational objectives, and working with him to formulate vocational plans. In plans and decisions of all these types, it is important to have a clear picture of the pupil's intellectual abilities as one aspect of the total picture of the pupil as an individual.

In educational guidance information about scholastic aptitude is especially important. This information should receive very serious consideration in deciding what is an appropriate educational objective for the pupil; i.e., whether to plan for college and if so the kind of college to plan for, or what type of high school curriculum to select. In vocational counseling, more specialized ability measures, of the kinds we shall consider in the next chapter, are desirable as a supplement to the general intelligence test, but these specialized tests are not so important for educational planning. For understanding a child who is having problems in school, whether with his school work or his personal adjustments, an estimate of his intellectual level is essential. As we have indicated elsewhere, individual tests and non-language tests are highly desirable supplements to the usual group test when any reading or language handicap is suspected.

The specific situations and circumstances under which intelligence tests may be used in guidance are so many and varied that they cannot each be discussed here. Some further consideration is given to tests in the guidance program in Chapter 18.

### SUMMARY STATEMENT

Tests of ability include tests of achievement and of aptitude. Though aptitude tests usually depend less directly upon specific teaching than do achievement tests, it must be recognized that any test performance is in some degree a function of the individual's background of experience. Aptitude tests are distinguished at least in part by their function—to predict future accomplishments.

Among the most thoroughly explored and widely used aptitude tests are tests of intelligence. As these have been developed, they tend to emphasize abstract intelligence, the ability to deal with ideas and symbols, and may even be thought of as scholastic aptitude tests.

The two main patterns of tests have been group tests and individual tests. Group tests, resembling the short-answer achievement test in format, are much more economical to use and are satisfactory for many purposes when the examinees are normal groups of school age or older.

However, the individual tests have a number of advantages and are useful particularly with (1) young children, (2) emotionally disturbed cases, and (3) cases with special educational disabilities.

Special tests have been developed for infant and preschool groups, for groups with educational and language handicaps, and for groups from varied cultures and social classes. These may be of practical value in special cases, though they serve more often as research tools.

Intelligence test results for school-age children are about as reliable as any of our psychological measurement tools. The widely used individual tests such as the *Stanford-Binet Intelligence Scale* and the *Wechsler Intelligence Scales* are probably somewhat more reliable than the typical group test, though the differences are not large. In spite of the high reliability, appreciable differences may be expected between one testing and another.

When intelligence test scores are studied in relation to achievement in the world, the most clear-cut relationships are with academic achievement. However, it is also true that there are substantial differences in test performance between persons in different types of jobs. Furthermore, success in at least some types of jobs has been found to be related to the abstract intelligence measured by our tests.

Group differences in intelligence (i.e., sex, race, age differences) must be interpreted quite tentatively, in view of the differences in background for these different groups. However, individual differences in intelligence are important facts, which we need to use wisely in helping individuals in their adjustment to the world of the school and of work.

## REFERENCES

1. Anderson, J. E., *The limitations of infant and preschool tests in the measurement of intelligence*, *J. Psychol.*, 8, 1939, 351-379.
2. Arthur, Grace, *A point scale of performance tests*, 2nd ed., New York, Commonwealth Fund, 1943.
3. Bayley, Nancy, Consistency and variability in the growth of intelligence from birth to eighteen years, *J. genet. Psychol.*, 75, 1949, 165-196.
4. Clark, W. W., *Questions and answers regarding the California Test of Mental Maturity*, Los Angeles, California Test Bureau, 1948.
5. Cornell, Ethel L., Effects of ability grouping determinable from published studies, in *The grouping of pupils*, *Nat. Soc. Study Educ.*, 35th Yrbk., Pt. I, 1936, 289-304.
6. Derner, G. F., M. Aborn, and A. H. Canter, The reliability of the Wechsler-Bellevue subtests and scales, *J. consult. Psychol.*, 14, 1950, 172-179.

7. Durost, W. N., and G. A. Prescott, An improved method of comparing a capacity measure and an achievement measure at the elementary school level, *Educ. Psychol. Meas.*, 12, 1952, 741-751.
8. Durrell, D. D., The influence of reading ability on intelligence measures, *J. educ. Psychol.*, 24, 1933, 412-416.
9. Ebert, E., and Katherine Simmons, The Brush Foundation study of child growth and development, I, Psychometric tests, *Monogr. Soc. Res. Child Developm.*, 8, No. 2, 1943,
10. Franzblau, R. N., Race differences in mental and physical traits, *Arch. Psychol.*, 1935, No. 177.
11. Gesell, A., et al., *The first five years of life: A guide to the study of the pre-school child*, New York, Harper, 1940.
12. Ghiselli, E. E., and C. W. Brown, The effectiveness of intelligence tests in the selection of workers, *J. appl. Psychol.*, 32, 1943, 575-580.
13. Goodenough, Florence L., *Measurement of intelligence by drawings*, Yonkers, N. Y., World Book, 1926.
14. Goodenough, Florence L., and Katherine M. Maurer, The mental growth of children from two to fourteen years; a study of the predictive value of the Minnesota Preschool Scales, *Univ. Minn. Inst. Child Welf. Monogr.*, No. 19, 1942.
15. Goodenough, Florence L., Katherine M. Maurer, and M. J. Van Wageningen, *Minnesota Preschool Scales: Manual of instructions*, Minneapolis, Minn., Educational Test Bureau, 1940.
16. Honzik, Marjorie P., Jean W. McFarlane, and Lucille Allen, The stability of mental test performance between two and eighteen years, *J. exp. Educ.*, 17, 1948, 309-324.
17. Justman, J., A comparison of the functioning of intellectually gifted children enrolled in special progress classes in the junior high school, unpublished doctor's dissertation, Columbia University, 1953.
18. Klineberg, O., *Negro intelligence and selective migration*, New York, Columbia University Press, 1935.
19. National Education Association, Research Division, Trends in city school organization, 1938 to 1948, *Res. Bull.*, 27, 1949, 4-39.
20. Raven, J. C., *Progressive matrices*, London, H. K. Lewis, 1956 (U. S. Distributor, Psychological Corp.).
21. St. John, C. W., Educational achievement in relation to intelligence as shown by teachers' marks, promotions and scores in standard tests in certain elementary grades, *Harvard Univ. Stud. Educ.*, 15, 1930.
22. Skodak, Marie, Children in foster homes: A study of mental development, *Univ. Ia. Stud. Child Welf.*, 16, No. 1, 1939.
23. Stewart, Naomi, A.G.C.T. scores of army personnel grouped by occupations, *Occupations*, 26, 1947, 5-41.
24. Stutsman, Rachel, *Mental measurement of pre-school children, with a guide for the administration of the Merrill-Palmer Scale of Mental Tests*, Yonkers, N. Y., World Book, 1931.
25. Terman, Lewis M., and Maud A. Merrill, *Stanford-Binet Intelligence Scale, Manual for the Third Revision, Form L-M*, Boston, Houghton Mifflin, 1960.

26. Thorndike, R. L., The prediction of intelligence at college entrance from earlier tests, *J. educ. Psychol.*, 38, 1947, 129-148.
27. Tuddenham, R. D., Soldier intelligence in World Wars I and II, *Amer. Psychologist*, 3, 1948, 54-56.
28. Wechsler, David, *Wechsler Adult Intelligence Scale*, New York, Psychological Corp., 1955.
29. Wechsler, David, *Wechsler Intelligence Scale for Children: Manual*, New York, Psychological Corp., 1949.
30. Wheeler, L. R., A comparative study of the intelligence of east Tennessee mountain children, *J. educ. Psychol.*, 33, 1942, 321-334.

### SUGGESTED ADDITIONAL READING

- Bayley, Naney, On the growth of intelligence, *Amer. psychologist*, 10, 1955, 805-818.
- Bradway, Katherine P., C. W. Thompson, and R. B. Cravens, Preschool IQ's after 25 years, *J. educ. psychol.*, 49, 1958, 278-281.
- Cronbach, Lee J., *Essentials of psychological testing*, 2nd ed., New York, Harper, 1960, Chapters 7 and 8.
- Dreger, Ralph Mason, and K. S. Miller, Comparative psychological studies of Negroes and Whites in the United States, *Psychol. Bull.*, 57, 5, 1960, 361-402.
- Eells, Kenneth Walter, et al., *Intelligence and cultural differences*, Chicago, University of Chicago Press, 1951.
- Froehlich, Clifford P., and K. B. Hoyt, *Guidance testing*, 3rd ed., Chicago, Science Research Associates, 1959, Chapter 5.
- Harris, Chester W., Editor, *Encyclopedia of educational research*, 3rd ed., New York, Macmillan, 1960, pp. 59-62, 715-717, 817-822.
- Miner, John B., *Intelligence in the United States*, New York, Springer, 1957.

### QUESTIONS FOR DISCUSSION

1. It has been proposed that all intelligence tests should really be called scholastic aptitude tests. What are the merits and the limitations of this proposal?
2. Why is it better to depend upon a good intelligence test for an estimate of a pupil's intelligence than upon ratings by teachers?
3. In each of the following situations would you elect to use a group intelligence test or an individual intelligence test? Why?
  - a. You are studying a boy with a serious speech impediment.
  - b. You are selecting students for a school of nursing.
  - c. You are preparing to counsel a high-school senior on his educational and vocational plans.
  - d. You are making a study of the Mexican children in a school system in Arizona.
  - e. You are working with a group of delinquents in a state institution.

4. In which of the following situations would you routinely first give the *Arthur Point Scale* rather than the *Stanford-Binet*? Why did you decide as you did?

- a. For testing Puerto Rican children entering school in New York City.
- b. For selecting children for a special class of gifted children.
- c. For evaluating intelligence in a school for the deaf.
- d. For studying children who have reading problems.

5. What are the implications for child placement agencies of the data on infant tests presented on p. 236?

6. Why do two different intelligence tests given to the same pupil quite frequently give two different IQ's?

7. Are the usual group intelligence tests more useful for guidance for professional occupations or for skilled occupations? Why?

8. A news article reported that a young woman who had been committed to a mental hospital with an IQ of 62 had been able to raise her IQ to 118 during the 3 years she had spent there. What is misleading about this news statement? What factors could account for the difference between the two IQ's?

9. In what respects are intelligence tests better than high-school grades as predictors of college success? In what respects are they less good?

10. Why do intelligence tests show higher correlations with standardized achievement tests than they do with school grades?

11. Comment on the statement: "College admissions officers should discount scholastic aptitude test scores of applicants who come from low socioeconomic groups."

12. You are a fourth-grade teacher. You have given a group intelligence test to your class and gotten IQ's from it. What additional information would you want to have on the pupils. What sorts of specific action and plans might grow out of the test results?

13. An eighth grader has received the following IQ's on the *Lorge-Thorndike Intelligence Test, Verbal*: Grade 4—98, Grade 6—112, Grade 8—102. What would be the best figure to represent his "true" scholastic aptitude?

14. A school in a prosperous community gave *Stanford-Binet* intelligence tests to all entering kindergartners and all first graders who had not been tested in kindergarten within the first week or two of school. How desirable and useful a procedure is this? Why?



## Chapter 10



# The Measurement of Special Aptitudes

The tests that we reviewed in Chapter 9 were tests of general mental ability. In most cases they resulted in a single score that represented an over-all appraisal of the individual's ability to deal with abstract ideas and relationships. However, we found that some of them did produce two or more scores of a more specialized nature that were designed to provide more specific and analytical information about the individual, i.e., the verbal and performance IQ's of the *Wechsler* scales. The concern for specific information on more restricted segments of the ability domain has led to the development of test batteries and single tests to measure specialized aptitudes. It is these tests that we shall consider in the present chapter. We will direct our attention first to batteries and tests designed for vocational guidance and vocational selection. Then we will consider specialized tests for prognosis and prediction in special school subjects and in special types of schools. Finally, we will take a brief look at tests in the specialized fields of art and music.

### VOCATIONAL APTITUDE BATTERIES AND TESTS

One of the early practical concerns of psychologists was in guiding young people into the types of work in which they would be happy and successful and in selecting for an employer those men who would be efficient and satisfied in the jobs that he was trying to fill. As psychologists began to study jobs, it seemed apparent that different ones required different special abilities as well as different levels of general mental ability. The automotive mechanic required a good deal of mechanical knowledge, but little verbal fluency, while the lawyer needed verbal comprehension but not mechanical skill. The book-keeper needed good ability with numbers, while the watchmaker needed fine coordination in his finger movements. The ability re-

quirements of jobs appeared to differ along a number of specialized dimensions.

At the same time, research demonstrated that human abilities are to some degree specialized. This has been shown in studies of the correlations between different tests. Consider the correlations shown in Table 10.1 between six tests of a battery used for classification of

Table 10.1. Intercorrelations of Selected Air Force Aptitude Tests

	1	2	3	4	5	6
1. Reading Comprehension	. .	.50	.05	.23	.13	.11
2. Navigator Information	.50	. .	.16	.25	.17	.15
3. Numerical Operations	.05	.16	. .	.44	.27	.11
4. Dial and Table Reading	.23	.25	.44	. .	.39	.23
5. Speed of Identification	.13	.17	.27	.39	. .	.43
6. Spatial Orientation	.11	.15	.11	.23	.43	. .

men in the U. S. Air Force.<sup>6</sup> Note that the correlations between the first two tests are relatively high. These are both tests that are quite verbal in nature and they appear to define a factor of ability to deal with verbal relationships. Tests 3 and 4 are both numerical tests and are substantially correlated. Tests 5 and 6, which correlate substantially with each other, both involve speed of visual perception. Note that the correlations of tests 1 and 2 with 3 through 6 are quite low. The verbal tests are measuring abilities quite different from those measured by the other four. The numerical and perceptual tests are not as clearly distinct from one another, but the correlations of tests 3 and 4 with 5 and 6 are less than the intercorrelation of 3 and 4 or the intercorrelation of 5 and 6. Thus, it appears that our six tests measure three somewhat distinct abilities: a verbal ability measured by 1 and 2, a numerical ability measured by 3 and 4, and a perceptual ability measured by 5 and 6. These abilities are not *entirely* independent but are tied together, perhaps by a common element of general mental ability running through all of them. However, the three are sufficiently different to justify separate measurement of them.

There has been a large volume of research on the organization and structure of human abilities during the last 50 years. Much of it has employed a technique known as *factor analysis* to try to tease out the underlying mental factors. Factor analysis starts with a table of correlations such as we have shown in Table 10.1 (usually, however, a much larger table) and tries to identify the pattern of underlying factors that could have produced the observed relationships. The tech-

niques are computationally laborious and statistically involved, and we shall not go into them in any detail here.\* We shall report merely that the research has indicated that one can distinguish quite a number of special ability factors, such as verbal comprehension, word fluency, numerical fluency, perceptual speed, mechanical knowledge, spatial visualizing, and inductive and deductive reasoning. It is also true that most of these abilities are to some degree related to each other. The tests of general intelligence discussed in the last chapter reflect a pooling of several of these separate factors, together with accentuation of their common core.

Through theoretical research on the nature of abilities on the one hand and the applied research on the validity of specific tests for specific jobs on the other, psychologists have been guided in the design of aptitude test batteries for use in educational and vocational guidance and in personnel selection and classification. Since about 1940, these batteries have come to occupy quite central positions in the testing scene, so we will need to study them in some detail. First, we will examine two of the most widely used batteries, one oriented primarily toward school use and the other toward industrial use. Then we will review some of the evidence on validity and consider the advantages and limitations of a battery of this sort.

*The Differential Aptitude Test Battery.* This battery was produced by the Psychological Corporation in 1947 as a guidance battery for use at the secondary-school level. Some attention was paid to getting measures of separate and relatively uncorrelated abilities, but the main attempt was to get measures that would be meaningful to high school counselors. As a result, the intercorrelations of the tests, with the exception of a test of clerical speed and accuracy, are about .50. However, the reliabilities of the separate tests average about .90 and are enough higher than the test intercorrelations to assure us that each test measures abilities somewhat distinct from those measured by the others. The eight subtests are briefly described and illustrated below.

1. *Verbal Reasoning.* Items are of the double-analogies type, i.e., ? is to A as B is to ?. Two sets of answer choices are provided and one must be picked from each set to complete the analogy.

*Example*

3. . . is to wide as thin is to .

1. store

2. narrow

3. nothing

4. street

A. fat

B. weight

C. man

D. present

\* For an introductory exposition of factor analysis see Guilford, J. P. *Psychometric Methods*. New York, McGraw-Hill Book Co., 1954.

2. *Numerical Ability.* Consists of numerical problems emphasizing comprehension rather than simple computational facility.

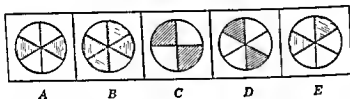
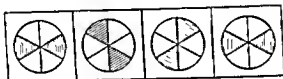
*Example*

$$\frac{1}{4} \div \frac{1}{8} =$$

A  $1\frac{1}{2}$   
 B  $\frac{3}{8}$   
 C  $\frac{1}{2}$   
 D 2  
 E none of these

3. *Abstract Reasoning.* A series of problem figures establishes a relationship or sequence, and the examinee must pick the choice that continues the series.

*Example*



A

B

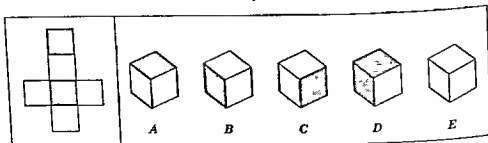
C

D

E

4. *Space Relations.* A diagram of a flat figure is shown. The examinee must visualize and indicate which solid figure or figures could be produced by folding the flat figure.

*Example*



A

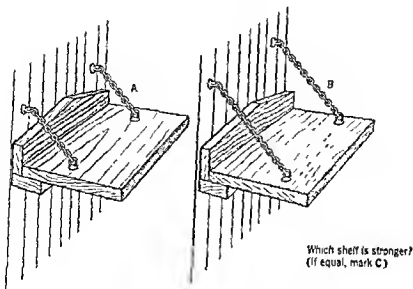
B

C

D

E

5. *Mechanical Reasoning.* A diagram of a mechanical device or situation is shown, and the examinee must indicate which choice is true of the situation.

*Example*

6. *Clerical Speed and Accuracy.* Each item is made up of a number of combinations of symbols, one of which is underlined. The examinee must mark the same combination on his answer sheet.

*Example*

Test Items

V.	<u>AB</u>	AC	AD	AE	AF
W	aA	aB	BA	Ba	<u>Bb</u>
X.	A7	7A	B7	<u>7B</u>	AB
Y.	Aa	Ba	<u>bA</u>	BA	bB
Z.	3A	3B	<u>33</u>	B3	BB

Sample of Answer Sheet

V	<u>AC</u>	<u>AE</u>	<u>AF</u>	<u>AB</u>	<u>AD</u>
W	BA	Ba	<u>Bb</u>	aA	aB
X	<u>7B</u>	B7	AB	7A	A7
Y	Aa	<u>bA</u>	bB	Ba	BA
Z	BB	3B	B3	3A	<u>33</u>

7. *Language Usage: Spelling.* A list of words is given, some of which are misspelled. The examinee must indicate for each word whether it is correctly or incorrectly spelled.

*Example*

	Right	Wrong
definate	<input checked="" type="checkbox"/>	<input type="checkbox"/>

8. *Language Usage: Sentences.* A sentence is given, containing one or more errors in usage or punctuation. The sentence is divided into subsections, and the examinee must indicate all the sections that contain an error.

*Example*

Ain't we/going to the/office/next week/at all.  
           A          B          C          D          E

*Sample of Answer Sheet*

A	B	C	D	E

The tests of the *DAT* are essentially power tests, with the exception of the Clerical Speed and Accuracy Test, and time limits are in most cases 30 minutes. Total testing time for the battery is about 5 to 5½ hours, and requires at least two separate testing sessions. Percentile norms are available for each grade from the eighth through the twelfth. Norms are provided for each of the subtests, and also for the combination of V and A, which may be used as a general appraisal of scholastic aptitude. An illustration of the profile form on which results may be plotted is shown on p. 152.

*The General Aptitude Test Battery (GATB).* The *General Aptitude Test Battery* was produced by the Bureau of Employment Security, U. S. Department of Labor, in the early 1940's. It was based upon previous work in which experimental test batteries had been prepared for each of a number of different jobs. Analysis of the more than 50 different tests that had been prepared for specific jobs indicated that there was a great deal of overlapping among certain ones of them, and that only about 10 different ability factors were measured by the complete set of tests. The *GATB* was developed to provide measures of these different factors. In its most recent form it includes 12 tests and gives scores for 9 different factors. One is a factor of general mental ability (G), resulting from scores on three tests (*Vocabulary*, *Arithmetic Reasoning*, and *Three-Dimensional Space*) that are also scored for more specialized factors. The other factors, and the tests that contribute to each are described below.

*Verbal Aptitude.* Score is based on one test, Number 4, *Vocabulary*. This test requires the subject to identify the pair of words in a set of four that are *either* synonyms *or* antonyms.

*Examples*

- |             |             |             |             |
|-------------|-------------|-------------|-------------|
| a. cautious | b. friendly | c. hostile  | d. remote   |
| a. hasten   | b. deprive  | c. expedite | d. disprove |

*Numerical Ability.* The appraisal of this aptitude is based upon two tests. The first of these, Number 2, *Computation*, involves speed and accuracy in simple computations with whole numbers.

*Examples*

Subtract (—)	256	Multiply (×)	37
	<u>83</u>		<u>8</u>

The second test entering into the *Numerical Ability* score, Number 6, *Arithmetic Reasoning*, involves verbally stated quantitative problems.

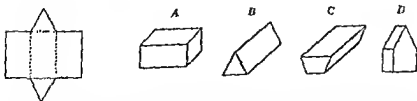
*Example*

John works for \$1.20 an hour. How much is his pay for a 35-hour week?

*Spatial Aptitude.* One test, Number 3, *Three-Dimensional Space*, enters into appraisal of this aptitude. The examinee must indicate which of four 3-dimensional figures can be produced by folding a flat sheet of specified shape, with creases at indicated points.

*Example*

Example of Spatial Aptitude

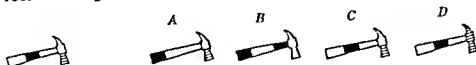


*Form Perception.* This aptitude involves rapid and accurate perception of visual forms and patterns. It is appraised in the *GATB* by two tests, Number 5, *Tool Matching*, and Number 7, *Form Matching*, which differ in the type of visual stimulus provided. Each requires

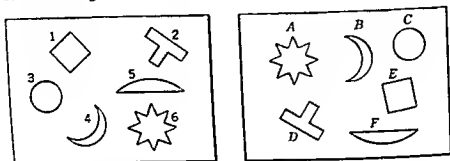
the examinee to find from among a set of answer choices the one that is identical with the stimulus form.

### Examples

#### Tool Matching:



#### Form Matching:



**Clerical Perception.** This aptitude also involves rapid and accurate perception, but in this case the stimulus material is linguistic instead of purely spatial. The test, Number 1, *Name Comparison*, presents pairs of names and requires the examinee to indicate whether the two members of the pair are identical, or whether they differ in some detail.

### Examples

John Goldstein & Co.—John Goldston & Co.  
Pewee Mfg. Co.—Pewee Mfg. Co.

**Motor Coordination.** This factor has to do with speed of simple but fairly precise motor response. It is evaluated by one test, Number 8, *Mark Making*. The task of the examinee is to make three pencil marks within each of a series of boxes on the answer sheet to yield a simple design. The result appears approximately as follows:



Score is the number of boxes correctly filled in a 60-second test period.



*Manual Dexterity.* This factor involves speed and accuracy of fairly gross hand movements. It is evaluated by two pegboard tests, Number 9, *Place*, and Number 10, *Turn*. In the first of these, the examinee uses both hands to move a series of pegs from one set of holes in a pegboard to another. In the second test, the examinee uses his preferred hand to pick a peg up from the board, rotate it through 180°, and reinsert the other end of the peg in the hole. Three trials are given for each of these tests, and score is the total number of pegs moved or turned.

*Finger Dexterity.* This factor represents a finer type of dexterity than that covered by the previous factor, calling for more precise finger manipulations. Two tests, Number 11, *Assemble*, and Number 12, *Disassemble*, use the same piece of equipment. This is a board with 50 holes in each of two sections. Each hole in one section is occupied by a small rivet. A stack of washers is piled on a spindle. During *Assemble*, the examinee picks up a rivet with one hand, a washer with the other, puts the washer on the rivet, and places the assembly in the corresponding hole in the unoccupied part of the board. He assembles as many rivets and washers as he can in 90 seconds. During *Disassemble*, he removes the assembly, returns the washer to its stack, and returns the rivet to its original place. Score is the number of items assembled or disassembled as the case may be. The apparatus tests are all arranged so that at the completion of testing the equipment has been returned to its original condition, and is ready for the testing of another person.

A comparison of the *GATB* and the *DAT* brings out that the *DAT* has tests of mechanical comprehension and language which the *GATB* lacks, while the *GATB* includes form perception and several types of motor tests that are missing in the *DAT*. Thus the *GATB* is more work oriented and less school oriented in its total coverage. Inclusion of the several types of motor tests results in somewhat lower correlations, on the average, for the *GATB*, though the "intellectual" tests correlate about as highly as those of the *DAT*. The correlations among the different aptitude scores of the *GATB* are shown in Table 10.2 for a group of 100 high school seniors. Excluding the correlations with *G*, which involves the same tests appearing in *V*, *N*, and *S*, the correlations range from  $-.06$  to  $.66$ . The three motor factors show fairly marked correlations, but they are practically unrelated to the remaining tests. The perceptual and intellectual tests also show quite a bit of relationship to one another, and this is most marked between the two types of perceptual tests.

Table 10.2. Intercorrelations of GATB Aptitude Scores for 100 High-School Seniors \*

	G	V	N	S	P	Q	K	F	M
G-Intelligence									
V-Verbal	73								
N-Numerical	74	42							
S-Spatial	70	40	34						
P-Form Percept.	43	34	42	48					
Q-Clerical Percept.	35	29	42	26	66				
K-Motor Coord.	-04	13	06	-03	29	29			
F-Finger Dext.	-05	-03	-03	01	27	20	37		
M-Manual Dext.	-06	06	01	-03	23	16	49	46	

\* Decimal points have been omitted.

There are quite substantial correlations between the corresponding factors of the *DAT* and the *GATB*. Representative values from one study\* are as follows:

Verbal	.70
Numerical	.56
Space	.69
Clerical	.56

However, the correlations are low enough so that it is clear that the tests cannot be considered identical. One important difference is the fact that the *DAT* tests are in most cases purely power tests, while the *GATB* tests are quite highly speeded.

*Other Aptitude Batteries.* A number of other aptitude batteries have been produced, mostly since 1950. There is generally less information available on these than on the *DAT* or the *GATB*, so their usefulness is less fully established. The batteries are briefly described in Appendix III.\*

There are also a good many single aptitude tests. Many of these are much like the tests that have been described as components of the *DAT* or *GATB*. The batteries have, of course, usually adapted ideas from the most effective single tests and incorporated measures that have been successful in previous use. Thus, the *Bennett Mechanical Comprehension Test* was the predecessor and model for the *DAT Mechanical Reasoning Test*. The *Minnesota Vocational Test* for

\* Fuller reports on each of seven different batteries, together with an evaluation by one outside expert, appeared in the *Personnel and Guidance Journal* from September 1956 through September 1957, and have been brought out as a separate monograph entitled *The Use of Multifactor Tests in Guidance*.

*Clerical Workers* provided the model for the *Clerical Perception* factor in the *GATB*. The various early mechanical aptitude and clerical tests have been reviewed by Bennett and Cruickshank,<sup>2,3</sup> and of course more recent tests will be found reviewed in the *Mental Measurements Yearbooks*.

#### VALIDITY OF APTITUDE BATTERIES

Now we must inquire into the usefulness of aptitude batteries such as the *DAT* and the *GATB*. We must inquire to what extent such a battery can provide us information that permits us to make better, more varied, and more differentiated predictions than those that are possible from a test of general mental ability or scholastic aptitude. The types of predictions with which we are most likely to be concerned are predictions of success in specific school subjects or major fields, predictions of success in specific jobs for which the individual is an applicant, and predictions of success in general fields of the world of work.

*Differential Prediction of Academic Success.* We have seen that scholastic aptitude tests have fairly good over-all validity for predicting academic success. One thing that we might hope is that an aptitude battery would tell us in which subject areas a student is most likely to be successful. Will Walter do better in English or in mathematics, in science or in French, in mechanical drawing or in history? A battery can do this to the extent that different tests in the battery are valid for different subjects. To what extent is this the case?

The manual for the *DAT* provides extensive data on the correlations of each of the subtests with achievement in a number of school subjects. Some of these results are summarized in Table 10.3. This

Table 10.3. Median Correlation of Differential Aptitude Test Scores with School Grades in Different Subjects

Test	English	Mathematics	Science	Social Studies, History	Language	Typing	Short-hand
Verbal Reasoning (VR)	.50(2)*	.39(1)	.54(1)	.50(1)	.37(1)	.19(6)	.44(3)
Numerical Ability (NA)	.45(3)	.60(1)	.51(2)	.44(2)	.42(1)	.37(1)	.27(4)
Abstract Reasoning (AR)	.35(5)	.35(4)	.44(4)	.35(5)	.25(3)	.27(3)	.24(3)
Space Relations (SR)	.27(8)	.32(3)	.25(6.5)	.26(6.3)	.13(8)	.15(7)	.16(6)
Mechanical Reasoning (MR)	.34(7.5)	.22(7)	.34(3)	.24(6)	.12(7)	.14(5)	.14(7.5)
Clerical Speed & Accuracy (CSA)	.31(7.5)	.19(8)	.26(8)	.26(6.5)	.23(1)	.20(4.5)	.14(7.5)
Spelling (Spelt)	.41(4)	.29(6)	.26(6.5)	.30(4)	.41(3)	.26(4.5)	.35(1)
Sentence (Sent)	.52(1)	.36(3)	.45(2)	.46(3)	.40(2)	.37(2)	.49(2)

\* Number in parentheses shows rank of that test for that subject.

table shows the median value of the correlations, and also ranks the eight subtests with respect to their correlations with each subject.

The first thing that we notice is that certain subtests are among the highest for almost all the subjects. Thus, *Verbal Reasoning* ranks near the top for all subjects except typing and *Numerical Ability* for all except shorthand. The *Sentences* test is one of the three most valid for all subject areas. This means that in large part the abilities that underlie academic performance are general abilities, and that a single general scholastic aptitude test will be effective in predicting success. The authors of the *DAT* have recognized this by printing the combined *Verbal Reasoning* and *Numerical Ability* tests as a single booklet and preparing separate norms for them. The combination of these two provides an effective measure of general scholastic aptitude.

At the same time, Table 10.3 does show some indication of differential validity. The *Mechanical Reasoning Test* is more valid for science than for the other subjects. The *Spelling Test* comes into its own in predicting success in shorthand. The *Numerical Ability Test* is more valid for mathematics than it is for English. A specific test does have a modest amount of differential validity, and does provide some suggestion that a pupil is likely to be more successful in one field than in another. However, it must be admitted that for much of educational guidance a general measure of scholastic aptitude will prove quite serviceable, and a battery of specialized aptitude tests will make only a limited additional contribution.

*Prediction of Specific Job Success.* We may next ask how successful a battery of aptitude tests will be in predicting the success of workers in a specific job in a specific company. Will the tests have validities high enough to make them useful to employers? Will different tests predict success in different jobs? The manual for the *GATB* provides quite an array of validities for job criteria. The data fall short of being ideal because the validation is often concurrent, based upon men already employed; because the samples are small; because the sample is typically limited to workers in a single plant or company; and because there is rarely any independent cross-validation.\* However, they provide about as good a pool of data as we have in which a common battery was validated against criteria of success in a number of different jobs. We have abstracted from the original report those instances in which validities are available against job (as distinct from school or training) criteria for samples of as many as seventy

\* Especially in exploratory studies in which a battery of tests is being tried out, it is important to verify validities discovered in an initial study by checking the same tests with a new independent sample.

cases and display them in Table 10.4. Only those correlations are shown in the table that are of a size that would be unlikely to have occurred by chance.\*

Table 10.4. Validity of GATB Scores for Specific Occupations

	Number of Cases	General Intelligence	Verbal	Numerical	Spatial	Form Perception	Clerical Perception	Motor Coordination	Finger Dexterity	Manual Dexterity
Bonding Mach. Op.	103		.21					.21	.23	.38
Bomb Fuse Parts Assembler	90		.34	.25	.43	.31	.23	.26	.37	.31
Chemist's Assistant	118	.30						.24		
Compositor, Hand and Machine	107	.39	.37	.40	.20		.34	.25	.50	.22
Laborer, Poultry	72	.24		.42			.25			.56
Machinist	71	.29			.37	.27	.30			
Moulder	281						.12	.22	.28	.38
Pottery Decorator	70				.25			.27		.31
Pressman, Cylinder	102	.40	.43	.52	.29	.39	.41	.27		.22
Sewing Mach. Op.	133	.27	.29	.36		.30	.26	.20		
Survey Worker	130	.50	.41	.44	.29					
Tabulating Machine Operator	203	.34	.22	.36	.20		.15			
Telephone Operator	88	.45	.38	.40	.27	.25	.35	.44	.23	
Underwriter	81			.24			.25			

From Table 10.4 we see that for every job two or more of the scores show significant validity. For some jobs all the validities are quite low, as for Pottery Decorator and Underwriter. For others, more encouraging values are obtained, as for Survey Worker. It is also clear that the factors that are valid for different jobs differ. Thus, manual dexterity appears to be important for a number of assembly and production line jobs, spatial ability counts relatively heavily for the machinist and chemist's assistant, clerical perception is relevant to the printing trades and to the underwriter, general intelligence discriminates the good from the poor survey worker, etc. In so far as these results are representative of the whole range of jobs, and within the limitations stated in the previous paragraph, they support the position that different specific abilities are valid for different jobs, and that a battery can be useful to an employer in picking men for a specific job or in assigning new workers to a type of assignment in which they will be effective and productive. Differences in success in certain jobs in single companies are predicted to a useful degree by an appropriate selection of aptitude tests.

A large number of separate studies of aptitude tests in relation to job

\* Correlations are exhibited that are statistically significant at the .05 level.

success have been summarized by Ghiselli.<sup>7</sup> Where a number of different sources provided correlations between scores on some type of test and success in a general category of job, he combined all the available data to produce a kind of pooled composite validity index. Selections from his report are shown in Table 10.5. Each entry is

Table 10.5. Selected Data on Average Validity of Different Sorts of Tests for Different Categories of Job (Adopted from Ghiselli)

Type of Test	Type of Job						
	Super- visory	Cleri- cal	Sales	Pro- tective	Vehicle Operator	Trades and Crafts	Indus- trial
Intelligence	28	31	02	27	14	20	20
Arithmetic	20	26	(06) *	(15)	(04)	23	13
Spatial Relations	21	10	...	(11)	...	19	14
Name Comparison	...	30	(-15)	(24)	...	(20)	16
Mechanical Principles	24	..	...	(27)	21	40	(50)
Finger Dexterity	...	24	...	(19)	...	20	18
Arm Dexterity	...	(18)	...	...	...	15	21

\* Correlations based on less than 500 cases are placed in parentheses.

an average, often of a number of correlations. The correlations have been enclosed in parentheses when they are based on less than five hundred persons. For some combinations of test and occupation no data could be found, so these entries have been left blank.

The pooled correlations reported by Ghiselli rarely go above .40, and then only for the smaller groups. Correlations in the twenties are fairly typical. For a given category of job, the variation in validity from one type of test to another is rather modest. Thus, these results present a rather less optimistic picture of the value of tests of special aptitudes than that portrayed in the *GATB* results in Table 10.4.

The less promising picture may stem in part from the blurring resulting from combining quite a span both of jobs and of tests within a single coefficient. It may be, however, that the larger numbers of cases represented in Ghiselli's composite correlations are less likely to

yield large correlations than the rather small U. S. Employment Service samples. The true picture of validity of tests as predictors of success at a given job in a given company, and of the distinctiveness of different abilities as predictors for different types of jobs probably lies somewhere between the pictures presented in these two tables.

*Forecasting Success in the World of Work.* For the school or college guidance counselor, special aptitude tests are useful in so far as they permit him to forecast some years in advance the general field of work for which a student will be able to complete training and in which he will be successful. The counselor cannot know what *specific* company the student will work for, or what exact job position he will fill. He deals in relatively long range forecasts over relatively broad categories. What evidence can be offered on the long-range forecasting effectiveness of aptitude test results?

Probably the most extensive study bearing on this problem is one in which approximately 10,000 men, who had originally been given an extensive battery of aptitude tests in the Air Force during World War II, were followed up some 13 years after the time of testing.<sup>14</sup> Test results were related to entry into and persistence in an occupation and to reported income and other indicators of success in that occupation. Even in a group of 10,000 men, samples in many occupations were small. However, it was possible to assemble samples large enough to merit analysis for about 125 occupational groupings.

The results on prediction of occupational success contrast rather sharply with those reported in the previous section. There was *no* convincing evidence of *any* relationship of test scores to success within an occupation for those men who had entered a specific occupation. Correlations were generally small, about as often negative as positive, and the total set of correlations could quite possibly have arisen as a result of chance deviations from a true correlation of zero. It appeared that when the men might enter an occupation such as law anywhere in the country, in many different kinds of settings both public and private, the test battery was quite unable to predict who would achieve the largest income, report the most satisfaction, or perceive himself as most successful in his field. It is important for the counselor to realize that such predictions are probably not possible for him.

For another type of occupational prediction the results were much more positive. Differences *between* occupations in average test score profile were real and quite marked for a number of occupations. Table 10.6 reproduces the results for selected occupations. Results for closely related tests have been combined into a factor score, and the table shows data for five distinct factors. Scores are standard

scores in which the mean for the complete population of aviation cadet applicants is set equal to zero and the standard deviation for this group is set equal to 100. Thus, a score of +50 represents a score half a standard deviation above the cadet applicant mean. The profiles of several of the occupations are shown graphically in Figure 10.1.

From Table 10.6 and Figure 10.1 we can see that there are substantial differences between one occupation and another. Most of these make good sense. The accountants as a group are highest on numerical ability, while the architects are highest on visual perception. Engineers are highest on the general intellectual measures that bulk so large in success in engineering school, while machinists are highest in mechanical and psychomotor skills. Some profiles have quite marked peaks and hollows, as, for example, the ones for accountant and machinist. Others are quite flat, exemplified by the sales engineer and

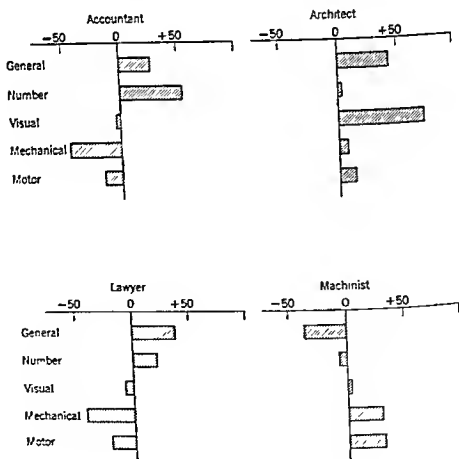


Fig. 10.1. Ability profiles for four occupations.



Table 10.6. Ability Profiles of Occupational Groups \*

Group	General Intellectual	Numerical Fluency	Visual Perception	Mechanical	Psychomotor
Accountants and auditors	28	54	- 4	-46	-16
Architects	44	4	74	8	14
Artists and designers	- 7	-12	51	- 4	8
Bricklayers	-24	- 5	-38	10	-32
Carpenters	-44	-17	- 4	24	- 1
College professors	75	38	38	-33	1
Contractors	- 7	-10	-10	34	5
Dentists	28	20	15	19	1
Draftsmen	1	-14	31	14	15
Drivers, bus and truck	-53	-11	-23	-14	-20
Engineers, chemical	106	42	30	19	20
Engineers, civil	75	31	56	36	14
Engineers, sales	57	33	35	39	40
Farmers, general	- 6	- 7	-29	38	-36
Lawyers	39	22	- 7	-42	-21
Machinists	-35	- 6	4	31	32
Mach. ops., fabricating	-45	-25	-25	-39	9
Managers, credit	- 5	22	25	-27	0
Managers, office	4	33	9	-29	11
Mechanics, vehicular	-72	-65	- 7	19	- 6
Physicians	59	20	18	2	0
Plumbers	-42	-21	-31	- 7	- 5

\* Expressed as standard scores: mean = 0, standard deviation = 100.

plumber profiles. These last two differ more in level than in any clear-cut patterning. Thus, we find some occupations with quite distinctive ability patterns, some with rather nondescript ones, some at a generally high level and some at a generally low level. Jobs vary noticeably both in the level of demand that they make and in their degree and type of specialization.

The differences shown in Table 10.6 and Figure 10.1 probably represent a minimum estimate of differences between occupations. This is because the Air Force group had already been screened by a test of general ability, so that the less intellectually able men who would ordinarily have been heavily represented in the semi-skilled and unskilled occupations had been screened out. On the other hand, it must always be remembered that although substantial differences can be shown *between* group means there is still a wide range of scores *within* each group. Some artists will be found with numerical ability higher than the typical accountant, some accountants with higher mechanical ability than the typical engineer. Differences between occupations are real, but so is variability within occupations.

### PROGNOSTIC TESTS

One group of aptitude tests is made up of tests designed to predict readiness to learn or probable degree of success in some specific subject or segment of education. These are called prognostic tests. A group of tests in this category that have been widely heralded and have received considerable use are the "reading readiness" tests. These tests are designed to be used with children, usually shortly after their entry into the first grade, to give the school as accurate an indication as possible of the child's ability to progress in reading. They provide information the teacher can use in assembling working groups within the class, in deciding upon the amount and type of prereading activities to provide, and in judging how soon to start a formal reading program. In some communities where kindergarten attendance is quite general, tests at the end of kindergarten are looked to as one basis for organizing first-grade groups for the following year. The sorts of tasks that appear in these tests may be seen from Table 10.7.

The reader who compares the tasks in Table 10.7 with the sample intelligence test items shown on pp. 222-225 will be aware of a substantial degree of similarity. In both, knowledge of word meanings appears. Both deal with recognition of sameness and differences, with analysis and classification. However, the reading readiness tests tend to emphasize more exclusively the materials of reading, letters and words. They include the components or early stages of the reading task. The basic question now becomes: Does the special slant which is given in the reading readiness test result in increased validity? Is the special test an improvement over a measure of general or academic aptitude? This is the question that must be raised for any type of prognostic test or special aptitude test.

Table 10.7. Types of Tasks Included in Representative Reading Readiness Tests

Type of Test Task	Gates	Lee-Clark	Metro-politan	Stevenson	Murphy-Durrell
Oral vocabulary or directions, using pictures	x	x	x		
Rhyming or matching sounds	x				x
Visual matching of figures, letters, or words	x	x	x	x	x
Visual perceiving of figures, letters, or words ("Which one is different?")		x		x	
Learning words in a standard lesson				x	x
Ability to read letters and words	x				

Whether a reading readiness test provides a better guide to later reading success than does a general intelligence test remains somewhat unclear. One fairly extensive investigation<sup>8</sup> indicates that tests requiring pupils to perceive and match words, to complete a story, and to select rhyming words gave better prediction of reading achievement one, two, or three terms later than did *Stanford-Binet* mental age. The validities reported for the *Gates Reading Readiness Test*, developed on the basis of this research, have been about .70, whereas *Stanford-Binet* MA showed a correlation of only .40 in the original study. This would indicate that the test tasks closely resembling the tasks faced by the beginning reader do have higher predictive effectiveness.

Another set of data<sup>10</sup> indicates, by contrast, that the *Pintner-Cunningham Intelligence Test* had a higher correlation with sixth-grade reading achievement than did the *Lee-Clark Reading Readiness Test*. However, these two sets of results need not be considered contradictory. The reading readiness test undertakes to predict ability to profit from reading instruction in the near future and is not used to forecast ultimate level of reading achievement. It may well be that it is more effective as an indicator of progress in reading within the next few months, even though an intelligence test is a better indicator of ultimate level of reading achievement.

Prognostic tests have been developed for various other subjects and levels, and the last few years have witnessed some renewal of interest

in these tests. Carroll and Sapon<sup>4</sup> have brought out a battery of foreign language prognosis tests and the *Symonds Foreign Language Prognosis Test* has been restandardized. The older Orleans prognostic tests for algebra, geometry and foreign languages continue to be available. The authors of all these tests offer evidence to show that the specialized tests provide a better prediction of achievement in the special subject area than is possible from a general measure of scholastic aptitude. However, one may still question whether, within the areas of academic achievement, special prognostic tests can improve the predictions based upon a combination of measures of general intelligence and previous academic achievement in related areas enough to justify their use. The demonstration that they can has not been sufficiently impressive to result in widespread adoption of the tests.

Special prognostic tests seem likely to be more useful as predictors of success in rather special types of academic tasks that have had no counterparts at earlier levels of school experience. Thus, the *Turse Shorthand Aptitude Test*, for which a correlation of .67 with later achievement in shorthand has been reported, may be useful as a supplement to other information about the pupil in evaluating probable success in shorthand training. The *ERC Stenographic Aptitude Test* and the *Bennett Stenographic Aptitude Tests* have given comparable results. These tests include such tasks as spelling, transcribing symbols, dictation under speed pressure, and word discrimination.

### PROFESSIONAL-SCHOOL APTITUDE BATTERIES

One other group of aptitude tests, so-called, are the tests that have been developed to select individuals for particular types of professional training. Many types of professional schools, sometimes individually but more often operating through their professional organizations, have instituted testing programs for the selection of their students. Testing programs are in operation for selecting students for engineering, law, medicine, dentistry, veterinary medicine, nursing, and accounting, to mention a few.

The tests used in these professional-school batteries tend to be tests of reading, quantitative reasoning, and apprehending abstract relationships, with the balance and emphasis shifted somewhat to conform to the academic emphasis of the particular training program. They are largely minor variations upon the same theme—a relatively high-level measure of scholastic aptitude and achievement. The different professional aptitude tests would correlate very substantially with one another or with a measure of general intelligence, and, indeed, it

should be expected that they would because the abilities required to succeed in training for the different professions have much in common. The similarities outweigh the differences. The common core is adapted to the professional field, as by giving more emphasis to quantitative materials for engineering and more to verbal materials for law. It is supplemented in some cases by rather highly specialized tests, for example, a test of chalk-carving for dentistry. These variations are superimposed upon the basic theme of scholastic aptitude and achievement.

## MEASUREMENT OF MUSICAL APTITUDE

\* When we come to such fields as music and art, the need for special measures of aptitude becomes quite apparent. Grades in these subjects are usually among those least well predicted by general measures of scholastic aptitude. Furthermore, the specialized nature of outstanding talent in these fields has long been recognized. Our problem is to determine what the components of this talent are and devise ways of appraising them.

In musical ability one large component is executive or motor, the ability to master the patterns of action required for playing an instrument. Aptitude measures have largely avoided this domain, perhaps because of its specificity to a particular instrument. Most measurement has been directed toward the perceptive and interpretive aspects of music.

Hearing music involves in the first place various types of sensory discrimination—discrimination of pitch, of loudness, of temporal relations. It involves in the second place perceiving the more complex musical relations in the material, interval relationships, the pattern of a melody, the composition of a chord, the relationship of a harmony to a melody. Third, it involves esthetic judgments about the suitability and pleasingness of a melody or harmony, a rhythmic pattern, or a pattern of dynamics.

The most thoroughly investigated musical aptitude test battery, the *Seashore Measures of Musical Talents*, is directed primarily toward measuring simple sensory discriminations, though with some attention to perceiving slightly more musical material. The tests have analyzed music down so far that very little music remains. Thus, there are the following subtests:

1. *Discrimination of Pitch*: judging which of two tones is higher.
2. *Discrimination of Loudness*: judging which of two sounds is louder.

3. *Discrimination of Time Interval*: judging which of two intervals is longer.

4. *Judgment of rhythm*: judging whether two rhythms are the same or different.

5. *Judgment of Timbre*: judging which of two tone qualities is more pleasing.

6. *Tonal Memory*: judging whether two melodies are the same or different.

The items are on phonograph records, with a series of items of each type. Within each type, the judgments become progressively more difficult.

The analytic approach to musical aptitude is evident in the above list of subtests. Critics have contended that the analysis has removed the tests a great way from any genuinely musical material and that fine discriminations of pitch, time, and intensity are really not called for in the activities of the musician. Validity studies of the Seashore tests have been somewhat conflicting, yielding appreciable correlations with measures of musical success in some instances and very low correlations in others. The value of the analytic test is still a matter of doubt and controversy.

Contrasting rather markedly with the Seashore type of test are the *Wing Standardised Tests of Musical Intelligence*. These tests, developed in England, were designed to stay as close as possible to the actual materials of music. The following subtests are included:

1. *Chord Analysis*: detecting the number of notes in a single chord.

2. *Pitch Change*: detecting the direction of change of one note in a repeated chord.

3. *Memory*: detecting which note is changed when a short melodic phrase is repeated.

4. *Rhythmic Accent*: judging which of two performances of the same piece has the better rhythmic pattern.

5. *Harmony*: judging which of two harmonies is more appropriate for a melody.

6. *Intensity*: judging which of two playings of the same piece has the more appropriate pattern of dynamics.

7. *Phrasing*: judging which of two renditions has the more appropriate phrasing.

This test is made up in part of tests that call for perceiving musical relationships and in part of tests that call for esthetic choices in intact musical material. Information on the validity of the test is still lim-

ited, but what there is seems very promising. Thus, the author reports<sup>12</sup> correlations of .64, .78, and .82 with teachers' rankings in three small samples. If these are maintained in future studies, a test like the *Wing* test would appear to have a very real place in guidance of young people who have musical aspirations or whose families hold such aspirations for them.

## TESTS OF ARTISTIC APTITUDE

Several types of tests are available relating to aptitude for art. In the first place, there have been tests of esthetic judgment. That field is now fairly well dominated by the *Meier Art Judgment Test*. Each item consists of a pair of pictures of art objects. One is an acknowledged masterpiece. The other is that same masterpiece systematically distorted in some specified way. The examinee must choose the better picture in each pair, the test blank indicating the respect in which the two specimens differ.

A test of the judgmental aspect of art ability is the *Graves Design Judgment Test*. This differs from the *Meier Test* in that all the items consist of abstract and non-representational material. The members of a pair differ in some single aspect of design, i.e., balance, symmetry, variety. Judgment of design is presumably divorced from any particular object or content.

In an attempt to get at the productive, as distinct from the purely judgmental, aspect of art, several tests (*Horn, Knauber, Lewerenz*) require the subject to produce drawings, based on certain limiting "givens." Thus, in the *Horn Art Aptitude Inventory*, a pattern of lines and dots is provided, and from this material the examinee must produce a sketch. The type of item is indicated in Fig. 10.2. The products must be evaluated by subjective rating, according to standards given by the authors, but they present some evidence that this can be done rather reliably even by non-artists.

The *Lewerenz Tests in Fundamental Abilities of Visual Art* use dot patterns to elicit drawings, whereas the *Knauber Art Ability Tests* use various assigned drawing tasks. Both these last two tests also present problems in shading, perspective, and composition.

Art tests have been rather generally successful in differentiating art students or art teachers from other groups. However, it has been argued that they accomplish this because they are in large measure achievement tests rather than aptitude measures. There has been relatively little study of these tests as aptitude measures with untrained individuals. Studies of art students have indicated that test perform-

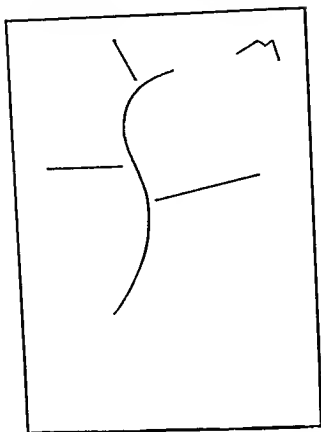


Fig. 10.2. Example of type of item used in Horn Art Aptitude Inventory.

ance is reasonably predictive of later art-school success. Thus, Horn and Smith<sup>2</sup> found a correlation of .66 between score on the Horn test at the beginning of the year and average faculty rating of success in a special high-school art class at the end of the year. Barrett<sup>1</sup> correlated four art tests with grades in a ninth-grade art course and with ratings of pupils' art products, with the following results:

	Course Grade	Ratings of Product
<i>McAdory Art Test</i>	.10	.13
<i>Meier Art Judgment Test</i>	.37	.35
<i>Knauber Art Ability Test</i>	.33	.71
<i>Lewerenz Fundamental Art Abilities Test</i>	.40	.76

Thus the last two tests, requiring production of drawings by the examinee, had about the same correlation with grades as did the *Meier Art Judgment Test* but much higher correlations with appraisals of student products.



We can see from the above that the test tasks that require art students to do the sorts of tasks they will be taught to do in art class predict their later achievement. How far down to untrained pupils this can be pushed remains to be determined.

Since the keying of art tests of all types depends upon a pooling of judgments, obtaining a high score requires conformity to the accepted esthetic standards. There is real question as to the applicability of these tests (or the tests of musical aptitude) in a distinctly different culture. There is also the possibility, though it is a fairly unlikely one, that a highly talented but unconventional person will be penalized on the tests.

### SUMMARY STATEMENT

Though general intelligence tests bear some relationship to success in many fields, efficient vocational guidance or personnel classification calls for tests more specifically directed at the abilities called for by each kind of job. Analytical studies of human abilities support the genuineness and importance of these special abilities. Numerous tests of special abilities have appeared, and more recently tests of this sort have been organized into comprehensive aptitude batteries for use in vocational guidance or personnel classification.

Special tests to evaluate readiness to undertake particular educational tasks have also been developed. The most widely used of these are reading readiness tests. Other types of prognostic tests have been less widely used, perhaps because their function is reasonably well served by measures of scholastic aptitude and academic achievement. Professional school aptitude batteries appear to be variations upon the basic theme of scholastic aptitude tests.

The fields of music and art have produced a number of ability tests. However, highly analytic tests have not been very clearly successful. More complex tests involve an unknown admixture of previous training. These show reasonably good validity and may provide an improved and at least relatively objective way of appraising status and, hence, promise in the field.

### REFERENCES

1. Barrett, H. O., An examination of certain standardized art tests to determine their relation to classroom achievement and to intelligence. *J. educ. Res.*, 42, 1949, 398-400.
2. Bennett, G. K., and Ruth M. Cruickshank. *A summary of clerical tests*. New York, Psychological Corp., 1948.

3. Bennett, G. K., and Ruth M. Cruickshank, *A summary of manual and mechanical ability tests*, New York, Psychological Corp., 1942.
4. Carroll, John B., and Stanley M. Sapon, *Modern language aptitude test*, New York, Psychological Corp., 1958.
5. DuBois, Philip, Editor, *The Classification Program*, Army Air Forces Aviation Psychology Program Report No. 2. Washington, D. C., U. S. Government Printing Office, 1947.
6. Gates, A. I., G. L. Bond, and D. H. Russell, *Methods of determining reading readiness*, New York, Teachers College, Columbia University, Bureau of Publications, 1939.
7. Ghiselli, Edwin E., *The measurement of occupational aptitude*, University of California Publications in Psychology, 1955, 8(2), pp. 101-216.
8. *Guide to the use of the general aptitude test battery, section III*, Bureau of Employment Security. U. S. Department of Labor, Washington, D. C., 1955.
9. Horn, C. A., and L. F. Smith, The Horn Art Aptitude Inventory, *J. appl. Psychol.*, 29, 1945, 350-355.
10. Lee, M. J., and W. W. Clark, *Lee-Clark Reading Readiness Test: Manual*, Los Angeles, Calif., California Test Bureau, 1951.
11. Thorndike, Robert L., and Elizabeth P. Hagen, *10,000 careers*, New York, John Wiley, 1959.
12. Wing, H., Tests of musical ability and appreciation: An investigation into the measurement, distribution, and development of musical capacity, *Brit. J. Psychol. Monogr. Suppl.*, 8, No. 27, 1948.

### SUGGESTED ADDITIONAL READING

- Froehlich, Clifford P., and K. B. Hoyt, *Guidance testing*, 3rd ed., Chicago, Science Research Associates, 1959, Chapter 6.
- Harris, Chester W., Editor, *Encyclopedia of educational research*, 3rd ed., New York, Macmillan, 1960, pp. 59-62, 1084-1085.
- Super, Donald E., *Appraising vocational fitness by means of psychological tests*, New York, Harper, 1949, Chapters 4, 6, 8-11, and 15.
- Super, Donald E., *The use of multifactor tests in guidance*, Washington, D. C., American Personnel and Guidance Association, 1958.

### QUESTIONS FOR DISCUSSION

1. A number of aptitude test batteries have been developed for use at the secondary-school level, but almost none for the elementary school. Why is this? Is it a reasonable state of affairs?
2. What are the advantages in using a battery such as the *Differential Aptitude Tests* instead of tests selected from a number of different sources? What are the limitations?
3. Step by step, what would need to be done to set up a program for selecting students for a dental school?
4. How could a high-school counselor use the data of Table 10.3? What are the limitations on the usefulness of this material?

5. How might the counselor use the data of Table 10.2? What are its limitations?

6. How sound is the statement, "The best measure of aptitude in any field is a measure of achievement in that field to date"? What are its limitations?

7. What are the differences between a reading readiness test and an intelligence test? What are the advantages of using the readiness test rather than an intelligence test for first-grade pupils?

8. To what extent are tests like the *Horn Art Test* and the *Wing Music Test* measures of aptitude? To what extent are they measures of achievement?

9. What factors tend to make tests of artistic and musical aptitude somewhat less useful than other types of aptitude tests?

10. In what ways could a follow-up study of graduates of a high school help in improving the school guidance program?

11. Why have aptitude test batteries shown up better in discriminating between jobs than in predicting success within a single job category?

## Chapter 11



# Achievement Tests

### STANDARDIZED VERSUS TEACHER-MADE TESTS

We turn now to tests of school achievement. We shall be concerned particularly with commercially available standardized tests, though we shall need to consider other types of appraisal devices in order to see how the standardized test fits into the total picture. As we indicated in the previous chapter, the distinction between an aptitude test and an achievement test is a somewhat blurred one. However, we shall be interested now in measures of knowledges and skills that are closely tied to organized school instruction, and in measures that are being used primarily to appraise present status in those school-taught knowledges and skills.

Standardized tests do not represent anything new and strange in the measurement of academic achievement. They are blood brothers of the short-answer teacher-made tests that were discussed in Chapter 4. They are made up of the same types of items and cover many of the same areas of knowledge. In what ways, then, do they differ from teacher-made tests? What are the advantages and limitations of each? For what purposes should each be used?

#### DISTINCTIVE FEATURES OF STANDARDIZED TESTS

There are four main ways in which commercially distributed standardized tests differ from the tests that the individual teacher would prepare for his own class.

1. The standardized test is based on the general content and objectives common to many schools the country over, whereas the teacher's own test can be adapted to content and objectives specific to his own class.
2. The standardized test deals with large segments of knowledge or skill, whereas a teacher-made test can be prepared in relation to any specific limited topic.
3. The standardized test is developed with the help of professional

2. Compare achievement of different skills or in different subject areas.
3. Evaluate the status of pupils from different schools or classes on a common basis, as when a pupil transfers to a new school.
4. Make comparisons between different classes and schools.
5. Study pupil growth over a period of time to see whether progress is more or less rapid than might be expected.

For some purposes, such as pupil diagnosis, we may wish to use not only standardized and teacher-made tests but a variety of informal testing and observational procedures as well.

We see, therefore, that standardized and teacher-made tests both have important functions to perform in the educational economy. To a large extent they are different functions. The two types of evaluation supplement one another. They are not competitors.

Standardized tests of achievement have been developed for practically every subject in the school curriculum. It would be impossible to give even a brief treatment of all subject areas in the pages that can be allotted to achievement tests in this book. We have decided, therefore, to concentrate on tests in a single area, in the belief that a fairly full treatment of this area will serve to introduce almost all the major problems and issues that would be encountered in dealing with tests in any area. We have chosen to discuss reading tests for two reasons. In the first place, these tests are more widely used than those in any other subject area. In the second place, a great variety of both survey and diagnostic procedures have been developed in this area, so that we shall get an introduction to a wide spectrum of testing techniques. Readers especially interested in tests in other areas can get names and evaluations of available tests in their field from *Buros' Mental Measurements Yearbooks* (see Chapter 8).

### ANALYZING THE OBJECTIVES OF AN AREA: THE FIRST STEP

Before we can proceed intelligently with either the preparation of a test in an area or the choice of one from among those already existing, we must analyze and define clearly the objectives of our instruction in the field in question.

Is test A a good test? Good for what? What are we trying to produce in children? What do we want or need to evaluate? Only when we have in our minds a clear answer to these further questions can we answer the question: Is this a good test? An analysis of objectives

helps us to identify the strengths and weaknesses of any single test or any complete evaluation program.

When we look at our statement of objectives, we will see some for which an existing test seems appropriate, some for which testing procedures might be devised if sufficient ingenuity were available, and some that seem obviously inaccessible by test procedures. Consider the following statement of objectives in the field of reading.\*

### Essential Knowledge, Attitudes, Skills, and Procedures in Reading

- I. Basic attitudes, skills, and procedures involved in securing meaning in both reading and listening.
  - A. To respond to the motive, problem, or purpose.
  - B. To direct attention to the meaning of what is read.
  - C. To develop fluent, accurate perception of word forms.
    1. Accurate discrimination of word forms.
    2. Accurate perception of both form and meaning.
    3. Association of right meanings with word forms.
    4. Accurate perception of words in context.
    5. Fluent perception of words.
  - D. To secure an adequate understanding of what is read.
    1. A clear grasp of meaning, involving:
      - a. Selection of meanings of words appropriate to the context.
      - b. Fusion of the meanings of words into a chain of related ideas.
      - c. Recognition of the importance and relationship of the ideas acquired.
    2. Coping successfully with such factors as:
      - a. Unusual word order.
      - b. Complexity of sentence structure.
      - c. Abstract ideas.
    3. Interpreting meaning in the light of its broader context. This ability implies an understanding of:
      - a. The total setting of the ideas expressed.
      - b. The author's mood, tone, and intention.
    4. Supplementing the specific meanings apprehended.
      - a. Reading between the lines.
      - b. Seeing implications.
  - E. To react critically to what is read.
    1. Recognizing the value, usefulness, timeliness, and significance of what is read.
    2. Judging the validity or truthfulness of the ideas presented in the passage.
    3. Judging the accuracy or completeness of the author's conclusions.

\* Adapted from Greene, Harry A., and William S. Gray, "The measurement of understanding in the language arts," *The Forty-Fifth Yearbook of the National Society for the Study of Education*, Part I, 1946.

4. Recognizing whether or not the reasoning of the author is sound.
5. Identifying and resolving propaganda.
- F. To integrate the ideas acquired with previous experience so that the following evidences of understanding may be noted immediately, or later:
  1. New insights are acquired.
  2. Previous understandings are reaffirmed or modified.
  3. Challenging problems are solved.
  4. Rational attitudes are acquired.
  5. Behavior is modified.
  6. Interests are broadened.
  7. Richer and more stable personalities are developed.
- II. Supplementary attitudes, skills, and procedures essential in many silent-reading activities.
  - A. To locate needed information.
    1. Using an index.
    2. Using a table of contents.
    3. Using the dictionary.
    4. Using card files.
    5. Using reference books.
  - B. To gather and evaluate information in the light of a given purpose.
    1. Recognizing the purpose to be achieved.
    2. Applying appropriate fact-finding techniques.
    3. Sorting essential from non-essential information.
    4. Judging the validity and significance of relevant information.
    5. Organizing the information in terms of the purpose or problem.
    6. Drawing tentative conclusions.
    7. Deciding when the purpose has been achieved.
  - C. To adjust reading attitudes and procedures to different purposes.
    1. Modifying interpretative processes in light of the purposes to be achieved. As, for example,
      - a. Reading to answer factual questions.
      - b. Reading an organized body of material to report.
      - c. Reading to determine the accuracy of the facts or events described.
    2. Adjusting rate of reading to the purpose.

As we scan this list of objectives, we see some that can obviously be measured quite readily and directly. Objective I C has to do with perceiving and comprehending words. Discrimination of word forms (I C 1) is appraised in tests for the primary grades by various types of word-picture or word-word matching tests. Accurate perception of form and meaning (I C 2) and association of form and meaning (I C 3) are central to the conventional vocabulary tests. One element entering into paragraph comprehension measures is accurate percep-

tion of words in context (I C 4). Fluent perception of words (I C 5) contributes to rapid reading and is appraised indirectly in a test of reading speed.

By the same token, most of the components under I D (to secure an adequate understanding of what is read) can be appraised by a well-designed test of comprehension of connected material. Thus, questions can be asked not only about the meaning of words in their context (I D 1a), but also about the sequence and relationship of parts of the passage (I D 1c) and about the author's mood or purpose (I D 3b). A teacher can test not only for facts explicitly stated, but also for implied conclusions (I D 4). A student can be called on not merely to comprehend the author's statement, but to judge the accuracy of his facts (I E 2), to appraise the soundness of a chain of reasoning (I E 4), or to identify the presence of propagandistic techniques (I E 5).

Though the better reading tests cover a wide range of skills of perceiving symbols, of getting meaning from them, and of evaluating and interpreting this meaning, no test can cover all of the objectives of reading. Thus, a reading test can hardly determine the pupil's responsiveness to reading (I A), or the extent to which he does in fact integrate reading into his total life experience (I F). Though certain components of study skills can be incorporated into a test (II A), a test will be less useful in appraising how the student actually uses these skills to solve an intellectual problem (II B).

Even more difficult might be the appraisal of the extent to which the individual *will* read and the type of reading he will select, as distinct from the extent to which he *can* read.

In this discussion we have been trying to make four main points. These points refer not only to reading but to any segment of the school program. The points are these:

1. The teaching of reading, or of any segment of the school program, is a complex undertaking looking to a variety of different outcomes.
2. A specific existing test will provide an appraisal of only certain ones of the desired outcomes.
3. Some of the desired outcomes are not likely to be reached by any test procedures.
4. The evaluation of any test of achievement requires a formulation of *your* objectives and an analysis of the test to see to what extent its content conforms to those outcomes that you are seeking to achieve.



## SURVEY READING TESTS

One of the most widely used types of standardized test is the survey reading test. Because of the importance of reading skills in all aspects of the school program, many schools make a special effort to appraise these skills as a basis for planning group activities and individual remedial action. As was suggested on pp. 291-292, reading is a complex and far-reaching enterprise. The commercial survey tests undertake to appraise only certain aspects of this range of skills. In Appendix III a number of the better-known and more widely distributed reading tests are listed, showing the grade levels for which forms are available, the types of subtests that are included in each, and certain other items of practical interest about each.

The subtests most frequently included in survey reading tests are word knowledge and paragraph reading. The test of paragraph reading usually involves paragraphs of some length with questions based upon each, though the pattern of a missing word or words to be supplied is sometimes used (*Stanford*) and the technique of requiring the reader to identify the word which *spoils* the meaning has also been tried (*Science Research Associates*). The paragraph with questions seems to correspond most naturally to the normal reading task.

The paragraph-with-questions pattern still leaves room for a wide range of variation in the processes that are actually tapped by the test. Only a critical examination of the single test items will enable the potential user to tell how many items call merely for knowing a word meaning, how many call for selecting the particular meaning which fits the context, how many for the answer to a specific factual question that is answered in the passage, how many for an inference based upon information given in the passage, how many for getting the main idea or theme of the passage, how many for sensing the author's mood or purpose, and how many for recognizing literary devices used in the passage. Reading with understanding is all these things and more. The different components are represented in different tests in very different proportions. The potential user of a reading test must examine the actual test items to get a real understanding of what abilities the test is measuring. And this is true not only for reading tests. In any area of achievement, it is only as the potential user examines critically the individual items on the test that he can judge whether this is a valid test for *his* purposes.

Fortunately, these different skills are all positively correlated, so that a survey based on some one combination of the skills will tend to

rank pupils in fairly much the same way as a survey based on a rather different balance of them. The child who does well on a test made up of directly factual items will in most cases tend to do well on one involving items of inference and synthesis. But the correlation is far from perfect. The potential user must examine each test in which he is interested, bearing in mind the specific types of comprehension skills that he deems important, in order to judge whether that particular test is the one best suited for his purposes. In the same way, a test in any subject matter must be scrutinized to see whether the test items represent the balance of information, understanding, and application that corresponds well with the objectives of teaching in the using school.

#### MEASURING SPEED OF READING

An additional factor that enters in to complicate the appraisal of reading achievement is the factor of speed. Speed of performance is a complicating factor in measuring achievement in any area, but perhaps especially so in the case of reading. To what extent do we want speed *per se* to enter into our score? Do we wish to penalize the person who is a slow worker but can accomplish a good deal if we give him time? Or do we want to get a pure measure of *power*, uncontaminated by speed of work?

Different tests resolve this problem in somewhat different ways, but generally speaking good testing practice accepts as its goal the separate measurement of *speed* of performance and *level* of performance. In measuring level, the objective is to provide enough time on any test so that each individual has had an opportunity to progress as far as his ability permits. This means either that he has had time to try all the items on the test or, if the items are graded in difficulty, that he has had time to work along to the point where he can no longer succeed with any of the test tasks.

In practical testing, this goal can only be approximated. Tests must be given with a definite time limit if they are to be fitted into a school program. Furthermore, it does not make for good testing conditions to have time limits so long that half the group is sitting around fidgeting. Time limits for a test designed to be a power test are usually figured so that *most* of the pupils have time to try *most* of the items.

But in the case of reading, we are also interested in speed for its own sake. Slow, laborious reading is inefficient and time-consuming. With some materials, it is desirable to be able to skim rapidly in order to cover a large amount of material in a short time. For this reason,

a number of tests have undertaken to include separate measures of speed of reading.

The measurement of speed presents its own special problems. We start a group of pupils reading an extended passage. At the end of 2 minutes we stop them and tell them to mark the word they were reading when the signal was given. But how do we know that they actually read the intervening material? Some may have read it word for word, others only skimmed, and others just read parts. What does "reading the passage" really mean?

Various devices have been tried in the attempt to make the task more uniform for different examinees. In some tests, such as the *lowo Silent Reading Tests* and the *Traxler High School Reading Test*, the reader is instructed to read in such a way that he will be able to answer questions. The *Gates Reading Survey for Grades 3 to 10* uses very brief paragraphs, each ending with a multiple-choice question, and score is the number of questions answered within the time limit. The *Michigan Speed of Reading Test* includes in each two-sentence unit one word that spoils the meaning of the unit. The reader must cross out these words. Any of these devices is only partly satisfactory. Any device that doctors up the actual text tends to distort the normal process of reading. Yet we cannot rely upon instructions alone to bring about comparable care and thoroughness by different readers. The reading speed test is at best very dependent upon the instructions given to the examinees and provides only an approximate indication of the relative speeds at which different people *can* read when degree of care and comprehension are uniform for each person.

#### SUMMARY STATEMENT

The survey test is, then, a sampling of the tasks that comprise a particular skill or knowledge. Only certain aspects of the total skill are represented, and the balance is a little different in each test. Most survey achievement tests have reasonably satisfactory reliability. The main problem is to pick the one that, in its content, provides the best balance of skills for your particular objectives.

#### DIAGNOSTIC TESTING

A survey achievement test undertakes to provide a general, over-all appraisal of status in some area of knowledge or skill. A diagnostic test undertakes to provide a detailed picture of strengths and weaknesses in an area. Furthermore, it is anticipated that this detailed analysis will suggest *causes* for deficiencies and provide a guide for

remedial procedures. A survey reading test tells us that Johnny, who is starting the fourth grade, performs on our test of reading paragraphs at a level typical of the usual child beginning the second grade. A series of diagnostic tests indicates that Johnny has a fair sight vocabulary of common words but no skills for working out unfamiliar words, that he is unable to blend sounds to form words, that he does not recognize the sounds that correspond to letter combinations, and that he makes frequent reversal errors. These findings, together with others, provide the basis for planning remedial teaching of word analysis and phonic skills that are specifically directed toward Johnny's deficiencies. Development of diagnostic tests involves two steps (1) analysis of the complex performance—be it reading, multiplying fractions, or using a microscope—into its component subskills and (2) developing tests for the component skills, free as far as possible from any other source of difficulty.

It has become fashionable in recent years to call many tests "diagnostic tests." In a sense, any test that yields more than a single overall score is diagnostic. Even if there are only two part scores, say, one for word knowledge and one for paragraph comprehension, the test makes it possible for us to say that Johnny showed better ability in word knowledge than he did in reading connected prose. This is certainly one diagnostic clue. Diagnosis is, after all, a matter of degree. We may probe and analyze with varying degrees of thoroughness and detail. We must ask concerning any test purporting to be diagnostic: How complete and how adequate are the diagnostic cues that this test provides? It is easy to overstate the value of the diagnostic information provided by a particular test.

Diagnostic testing faces a very troublesome dilemma. How is the test to provide sufficient diagnostic detail and yet appraise each separate ability with sufficient reliability? The essence of good diagnosis is that one should get many distinct and relevant facts about the individual. One wishes an appraisal of each of the component abilities into which the complex performance has been analyzed. At the same time, it is important that the separate appraisals have adequate reliability.

Reliable appraisal is particularly important in diagnostic testing for two reasons. In the first place, in diagnostic work we are in almost every instance interested in the *individual*. It is his personal strengths and weaknesses with which we are concerned. Group averages or group comparisons are of no particular interest to us in this context. We cannot fall back upon averages to balance out the chance errors in measuring a particular pupil. We need an accurate appraisal of

the specific individual. This is made more acute by the fact that we are dealing with *differences* between the individual's performance in related tasks. We are interested in making such a statement as: "This pupil's ability to pick out the main idea in what he has read is poorer than his ability to answer questions on specific factual details." But the two abilities are quite closely related. How reliably can we measure the differences between the two?

At this point, the student could well refer to the discussion of profiles on pp. 147-152. A set of diagnostic scores is a specific instance of a profile. All the issues about the reliability of a difference score that were raised in that discussion apply very acutely to the case of diagnostic tests. Since we are dealing with different aspects of a single field, correlations between tests are likely to be fairly substantial and the loss of reliability to be considerable when we have to think about differences. One would think, this being so, that authors of diagnostic tests would have been particularly concerned about the reliability of their instruments. But, alas, this has not generally been the case. The temperament that becomes excited about problems of diagnosis appears to be different from the temperament that grows concerned about issues of reliability of measurement. It must be confessed that in many cases the reliability of diagnostic tests is quite modest and that in many others it is unreported.

All this means that diagnostic test results must be interpreted with caution. The tests provide some rough and quite tentative hypotheses as to the individual's strengths and weaknesses. But these must be clearly recognized as tentative hypotheses and nothing more. The test profile suggests possible causes for the present difficulty and a jumping-off place for remedial work. If the remedial activities are successful, well and good. If not, the remedial teacher must stand ready to review his hypotheses and to explore other leads. Diagnostic test results are suggestions, not commands.

We find several types of diagnostic instruments in reading, and these serve also to illustrate the varieties of diagnosis in other areas. In the first place, we find tests with somewhat specialized subtests yielding scores for some aspect of the total function. This type is well illustrated by the *Iowa Silent Reading Tests (Advanced Level)*. These have the following subtests, each supposed to represent a somewhat different aspect of reading skill.

Test 1. *Rate and Comprehension* of connected prose.

Test 2. *Directed Reading* of connected prose to locate answers to factual questions.

Test 3. *Poetry Comprehension*, including mood, metaphor, etc.

Test 4. *Word Meaning* in different content areas.

Test 5. *Sentence Meaning* of brief sentences out of context.

Test 6. *Paragraph Comprehension*: selecting central idea and comprehending essential details

Test 7. *Location of Information*: using an index, selecting key words.

How many of these are in fact both sufficiently reliable and sufficiently different to be usefully diagnostic is a real question. For example, the reliabilities of *Test 5, Sentence Meaning*, and *Test 6, Paragraph Comprehension*, are reported (probably somewhat optimistically, since the coefficients are based on odd versus even halves and the tests have quite short time limits) as .751 and .759. The correlation between the two tests is reported as .48. From these values, we may estimate the reliability of the difference score to be .53. Inferences from a datum having this level of reliability should be made very cautiously.

The use of subtest scores such as those on the *Iowa* is probably most justifiable for a class or larger group. With a group average, chance errors tend to cancel out, and the low reliability of the scores becomes less important. If the group as a whole shows some marked weakness, as in the use of indices and library aids, for example, this may point out areas in which instruction has been neglected and suggest directions for instruction for the group as a whole.

A second approach to the diagnostic study of reading is through standard oral reading passages. One test of this type that has been used for many years is *Gray's Oral Reading Passages*. The test consists of a standard set of passages, ranging from easy and simple to quite difficult. The child who is being studied reads the passages aloud. The examiner uses a standard code to record on a copy of the passages all the errors and hesitations made by the pupil. Mispronounced words are underlined. Mispronounced vowels are shown by appropriate diacritical marks. Omissions are encircled. Substitutions and insertions are written in. Repetitions are indicated by a wavy line. A sample record with a number of errors indicated upon it is shown in Fig. 11.1.

The sun pierced into <sup>may</sup> my large windows. It was the opening of October, and <sup>clear</sup> the sky was <sup>of a</sup> of a dazzling blue. I looked out of my window and down the street. The white house of the long, straight street were on most painful to the eyes. The clear atmosphere allowed full play to the sun's brightness.

Fig. 11.1. Example of reading passage taken from *Gray's Oral Reading Passages*. (Reproduced by permission of the Public School Publishing Co.)

The record of the child's oral responses is valuable for the insight that it gives us into the actual *process* of reading. The usual objective written test shows us only the *product* of a child's efforts, the marks he makes on a test booklet or answer sheet. If he does poorly or makes mistakes, we are often at a loss to know why. In the oral test we can see the errors as they happen—each hesitation, each omission, each reversal. In this way we can identify more specifically the components that are giving the child trouble. They are not lost in the one final result, that the child is slow in reading the passage or does poorly on comprehension questions based on it.

The oral test as a basis for diagnosis can be illustrated in arithmetic also by the *Buswell-John Diagnostic Test for Fundamental Processes in Arithmetic*. This test consists of a series of graded examples. The examples are to be worked out by the child "thinking out loud," telling what he is doing and why he is doing it at each step. The examiner has a record sheet, with a code for types of erroneous processes. One page of the record sheet is illustrated in Fig. 11.2. The examiner uses this form to record errors made by the pupil as he speaks out his solution of the problem. A study of the types of errors that the pupil is making may suggest specific points at which the pupil needs help. This opportunity to gain insight into the way in which the pupil is attacking the task and to understand the nature of his errors is an advantage of oral testing procedures in whatever field they may be used.

In a third type of diagnostic test the test maker tries to analyze the complex task, such as reading, into its simpler components and test these components one at a time. Thus, the *Gates Reading Diagnosis Tests* include tests of recognition of words, recognition of separate syllables, ability to blend the sounds of letter combinations, and recognition of the single letters. The complex skill is pushed back to smaller and smaller segments of the total task. The thought is that when a person fails on the complex task we test to see whether he is able to show the component skills of which the larger task is built.

This type of approach may be illustrated in another field by the *Compass Diagnostic Arithmetic Tests*. In these the authors undertake to break up each complex skill in arithmetic into its components—to test the simplest components first, and then to add on additional elements until the full task has been tested. Thus, the diagnostic test concerned with division of whole numbers has subsections testing the child upon the following contributing skills and understandings: (1) the vocabulary of division, (2) fundamentals of short division, (3) short division with carrying, (4) the addition, subtraction, and multiplication used in later subtests, (5) estimating the first quotient fig-

**DIAGNOSTIC CHART**  
FOR  
**INDIVIDUAL DIFFICULTIES**  
**FUNDAMENTAL PROCESSES IN ARITHMETIC**  
*Prepared by G. T. Russell and Louise Jahn*

Published by the  
Public School Publishing Co.  
Bloomington, Illinois

Teacher's Diagnosis \_\_\_\_\_  
for Pupils \_\_\_\_\_

Name \_\_\_\_\_ School \_\_\_\_\_ Grade \_\_\_\_\_ Age \_\_\_\_\_ IQ \_\_\_\_\_

Date of Diagnosis \_\_\_\_\_ Add \_\_\_\_\_ Sub \_\_\_\_\_ Mult \_\_\_\_\_ Div \_\_\_\_\_

Teacher's preliminary diagnosis \_\_\_\_\_

**ADDITION:** (Place a check before each habit observed in the pupil's work)

<ul style="list-style-type: none"> <li>— a1 Errors in combinations</li> <li>— a2 Counting</li> <li>— a3 Added carried number last</li> <li>— a4 Forgot to add carried number</li> <li>— a5 Repeated work after partly done</li> <li>— a6 Added carried number irregularly</li> <li>— a7 Wrote number to be carried</li> <li>— a8 Irregular procedure in column</li> <li>— a9 Carried wrong number</li> <li>— a10 Grouped two or more numbers</li> <li>— a11 Split numbers into parts</li> <li>— a12 Used wrong fundamental operation</li> <li>— a13 Lost place in column</li> <li>— a14 Depended on visualization</li> </ul>	<ul style="list-style-type: none"> <li>— a15 Disregarded column position</li> <li>— a16 Omitted one or more digits</li> <li>— a17 Errors in reading numbers</li> <li>— a18 Dropped back one or more tens</li> <li>— a19 Derived unknown combination from familiar one</li> <li>— a20 Disregarded one column</li> <li>— a21 Error in writing answer</li> <li>— a22 Skipped one or more decades</li> <li>— a23 Carrying when there was nothing to carry</li> <li>— a24 Used scratch paper</li> <li>— a25 Added in pairs, giving last sum as answer</li> <li>— a26 Added same digit in two columns</li> <li>— a27 Wrote carried number in answer</li> <li>— a28 Added same number twice</li> </ul>
--	---

Habits not listed above \_\_\_\_\_

(Write observations under the pupil's work in space opposite examples)

(1)	$\begin{array}{r} 8 \\ 3 \\ \hline \end{array}$	$\begin{array}{r} 6 \\ 2 \\ \hline \end{array}$	(5)	$\begin{array}{r} 9 + 2 = \\ 3 + 4 = \end{array}$		
(2)	$\begin{array}{r} 3 \\ 6 \\ \hline \end{array}$	$\begin{array}{r} 8 \\ 4 \\ \hline \end{array}$	(6)	$\begin{array}{r} 22 \\ 12 \\ \hline \end{array}$	$\begin{array}{r} 40 \\ 38 \\ \hline \end{array}$	
(3)	$\begin{array}{r} 12 \\ 2 \\ \hline \end{array}$	$\begin{array}{r} 13 \\ 8 \\ \hline \end{array}$	(7)	$\begin{array}{r} 29 \\ 21 \\ \hline \end{array}$	$\begin{array}{r} 48 \\ 38 \\ \hline \end{array}$	
(4)	$\begin{array}{r} 19 \\ 2 \\ \hline \end{array}$	$\begin{array}{r} 17 \\ 8 \\ \hline \end{array}$	(8)	$\begin{array}{r} 3 \\ 8 \\ 6 \\ 2 \\ \hline \end{array}$	$\begin{array}{r} 8 \\ 7 \\ 8 \\ 7 \\ \hline \end{array}$	

Fig. 11.2. Example of record sheet used in the Buswell-John Diagnostic Test for Fundamental Processes in Arithmetic. (Reproduced by permission of the Public School Publishing Co.)



ure, (6) fundamentals of long division and checking, and (7) finding errors in long division. A study of scores on these subsections may provide insight as to where the trouble *really* lies.

Related to this type of test is the test that is loaded with opportunities to make a particular type of error. Thus, one test used by Gates in reading diagnosis is one in which the examinee reads a set of words that lend themselves to reversal errors, i.e., was—saw, on—no. Such a test gives a concentrated exposure and permits a judgment of the susceptibility of the examinee to that particular type of error. Informal tests of this sort in such fields as language usage, spelling, etc., are, of course, familiar to any teacher who tries to check upon the effectiveness of his teaching of particular usages, rules, generalizations, and understandings.

Finally, diagnostic testing in any given field must go beyond the immediate field of skill or knowledge and seek information on all the background factors that contribute to success or difficulty in the particular area. Thus, to understand the child with reading difficulty we need information on his vision, his hearing, his general intellectual level, even his interests and his emotional adjustment. So a thorough diagnostic study will include tests of visual acuity, muscular balance and fusion, testing with an audiometer to be sure the child can hear adequately, a non-reading intelligence test, and interview or questionnaire information about factors in the child's background and present life that may prove relevant. Diagnostic testing spreads out beyond subject boundaries and a full diagnostic study becomes essentially a directed case history of an individual, directed in that it is focused on the academic problem but comprehensive in that it covers all potentially significant features of both the skill area and the individual's personal life.

We have described a variety of diagnostic procedures in reading and in arithmetic. It is in these fields that the most work on diagnostic procedures has been done. There are, in fact, few published diagnostic tests outside these fields, though there is certainly informal teacher diagnosis. Even in the fields of reading and arithmetic, relatively little information about the specific diagnostic tests is provided by the authors. Evidence on reliability is meager, and norms are rather crude and fragmentary. Most diagnostic tests are not very elegant psychometric devices. They have not been sufficiently widely used to support the large investment in development and analysis that characterizes the more popular survey tests. Interpretation of test scores must, therefore, be made with particular care and a good deal of tentativeness.

## ACHIEVEMENT MEASURED THROUGH PUPIL PRODUCTS

One type of achievement measure that cannot be well illustrated within the field of reading is the product scale. We can illustrate this type of appraisal in the field of handwriting. Here, the plan is to evaluate some performance of an individual, in this case his handwriting, by comparing it with a set of standard samples. The standard samples are chosen by using the pooled judgments of a number of judges. The judges are usually asked to consider specimens in pairs and decide which is better. The basic idea in this type of scaling is that the larger the per cent of judges who agree in noticing a difference, the larger the difference. Thus, if 90 per cent of judges consider specimen A to be better than specimen B and only 80 per cent consider B to be better than C, the difference between A and B is greater than the difference between B and C. If 50 per cent consider C better than D and 50 per cent consider D better than C, then C and D must be considered of equal merit. Equally perceptible differences are considered to be the same size. Thus, a difference that is agreed upon by 75 per cent of our group of judges would be considered to be the same size wherever it occurred on our scale. Basing our scale units on this case-of-perception standard, we can set up a scale of specimens from very poor to very good and assign a numerical value to each.

When we use a product scale, the procedure is to compare the specimen of a pupil's performance with the set of standard samples. His product is moved up and down the set of standard samples until the judge decides which one it most nearly resembles. It is then assigned the scale value of the one that it matches most closely. If greater accuracy is desired, each specimen may be compared to the set of standard samples by two or more judges independently and their judgments averaged to give the final value.

Product scales have been used for such performances as handwriting, sewing, drawing, and manual arts. They are potentially applicable to any area of skill in which a permanent tangible product is the end result.

## ACHIEVEMENT TEST BATTERIES

The tests that are probably most widely used in programs of achievement testing are survey achievement test batteries. These batteries represent "package" achievement testing programs ready-made for the

schools' use. The typical battery is made up of from four to eight or ten separate tests covering the core knowledge and skill segments of the curriculum. We shall examine the content of several batteries in more detail presently. The attempt of the authors and publishers is to produce an integrated instrument that will cover the general achievement testing needs of the typical community.

The chief virtues of the single battery of tests, as compared with a program made up of separate tests chosen from a variety of different sources, are those of unity and of convenience. A test battery is unified in two important respects. In the first place, it is based upon a unified and integrated plan. The parts have been selected and the content of each planned with an eye to the whole. Within the limits of the professional skill and understanding of the team of authors, the product is a unified whole in which the parts fit together to cover the range of objectives that they deem important and feasible to appraise with a standardized test.

A battery is unified in one other important respect. It has a unified set of norms. The norms for all the subtests are based upon the same population and expressed in the same form. This makes direct comparison between the different subtests possible with a minimum of question. We do not have to ask whether our reading test was tried out on the same type of group as our arithmetic test, or how the standard scores of our spelling test compare with the percentile equivalents of our language usage measure. When tests are assembled from different sources, these problems can be matters of real concern. Particularly in the past, when norming populations for tests were assembled in a somewhat haphazard manner, the comparability of a grade score of 4.0, for example, from one test to another was subject to serious question. The large, broadly representative groups used in norming recent achievement batteries assure both breadth of representation for the norms as a whole and equivalence of meaning from one test to another.

Of course, the "package" testing program based on a standard battery has certain limitations. The chief one is rigidity. Some sections of a battery may fit a particular local curriculum better than others. Some subtests of one battery may fit modern curricular objectives, whereas another battery may seem better in another area. The user of the battery gets the good with the bad, "the bitter with the sweet." Short of omitting certain sections completely, he must use what the battery offers him, even though in certain respects it may not fit his needs, as he sees them, as well as some other specific test covering that

area. How serious this is the consumer must judge for himself when he compares the subtests of the battery that he is using or proposes to use test by test with other tests that are available for measurement in those same areas. The general verdict of users, particularly in the elementary school, has been that the convenient and unified program represented in a survey battery has more advantages than drawbacks, and in practice such instruments are very widely used.

#### COMPARISON OF ELEMENTARY SCHOOL BATTERIES

We propose to try to give a picture of the common features of some of the widely used achievement batteries. Since content changes somewhat at different levels, and space does not permit a comparison at all levels, we have chosen the tests designed for use at about the fifth and sixth grades. The upper elementary school is probably the level at which achievement test batteries are most widely used. We have elected to compare the following batteries, with publishers and approximate publication dates as indicated:

- California Achievement Tests*, California Test Service, 1957
- Iowa Tests of Basic Skills*, Houghton Mifflin Co., 1956
- Metropolitan Achievement Test*, World Book Co., 1960
- Sequential Tests of Educational Progress (STEP)*, Educational Testing Service, 1957
- SRA Achievement Test*, Science Research Associates, 1956
- Stanford Achievement Test*, World Book Co., 1956

The tests will be compared area by area, to bring out the elements that are common and the features that are distinctive.

#### MEASUREMENT OF WORD KNOWLEDGE

Each of the tests except the *STEP* provides for the appraisal of word knowledge. However, the tests vary in the degree to which this ability is kept separate for scrutiny as a significant fact about the individual. On the one hand, the *SRA* tests appraise vocabulary only in paragraph context, and include the items only as part of a total appraisal of reading ability. By contrast, the *Iowa* and *Metropolitan* tests yield a separate vocabulary score, and provide no procedure for putting it together with paragraph reading in a single reading score. The others (*California*, *Stanford*) provide a separate word knowledge score, but also provide for combining this with paragraph reading in a total reading score.

## MEASUREMENT OF READING

Every one of the tests provides for the measurement of reading ability as represented by the reading of connected passages. The tests vary quite widely, however, in the length of the passages, and the type and range of test items based on each. At one extreme, appraisal in the *Stanford Achievement Test* is based on passages only 50 to 100 words long with two or three items on each passage. (These follow the somewhat unusual format of omitting words or phrases from the passage and requiring the subject to pick the word or phrase that best fills in the gap—an activity rather different from the normal process of reading.) At the other extreme, the reading test of the *SRA Achievement Series* is based upon a small number of long passages (500 or 600 words) with as many as twenty test items referring to a single passage. The other tests use passages of intermediate length with six to ten comprehension items testing various aspects of the comprehension of each passage.

As noted above, a number of the tests combine word knowledge with paragraph comprehension into a single global reading score. One may question whether it is desirable to have knowledge of words *per se* bulk so large in an appraisal of reading. Of course, word knowledge is quickly and easily measured, but real reading comprehension would seem to be better exemplified by ability to get meaning and draw inferences from connected material.

## MEASUREMENT OF ARITHMETICAL SKILLS AND UNDERSTANDINGS

All achievement batteries make some appraisal of ability in arithmetic. The older batteries tended to break the total area of arithmetic up into computational skills and problem solving, and to provide two subtests corresponding to these two areas. Provision was usually made for combining the subtests into a single global appraisal of arithmetical ability. Among many of the newer tests, however, there is an additional concern with arithmetical concepts and understandings. In the *Iowa tests*, the computational subtest has been entirely replaced by a test dealing with arithmetical concepts. In others (*Metropolitan*, *SRA*) a section dealing with concepts is included in either the skills or the problem-solving subtest. Addition of the material on concepts reflects the increased emphasis within the arithmetic curriculum on developing meaning and understanding, as distinct from simply facility in carrying out mechanical operations.

In arithmetic it is often difficult to free appraisal of problem-solving ability from the influence of reading skills. This is illustrated by the

mathematics test of the *STEP*. In a commendable attempt to incorporate the arithmetical tasks in real and meaningful problems, the authors have introduced a reading level such that the resulting score has a higher correlation with a verbal than a quantitative score on the parallel aptitude series (*SCAT*).

#### LANGUAGE SKILLS

Another common denominator in all the batteries is appraisal of various skills in using language. The batteries vary in detail, but typically they cover capitalization, punctuation, elements of usage such as case, number and tense, and spelling. As in other subtests, there is a tendency here too to present all tasks in multiple choice form. Thus, most of the spelling tests call in some manner for the recognition of error. The examinee must decide which word in a set, if any, is misspelled (*Iowa*); or he must decide whether or not a given word is misspelled, and if it is he must correct it (*Metropolitan*). Recognition of the correct form is also typical of the usage items, and it is assumed that the ability to recognize error is a good indicator of ability of the student to avoid error in his own writing.

The psychometrician is fairly happy with this assumption, on the basis of the correlation of recognition test scores with ratings of quality of actual writing, and is impressed by the greater reliability, efficiency and objectivity of the recognition test. The English teacher, however, would still prefer to appraise writing ability through an actual sample of writing—an essay of some sort. The only one of the achievement batteries that provides for this is the *STEP*. The essay test is supported in part because it may measure, though unreliably, something that is not reached by the objective tests. It is justified further as having a healthy effect on the curriculum, the argument being that we must evaluate writing if we are to expect the schools to continue to teach it.

One further aspect of language skill that appears in the *STEP* is a test of listening comprehension. Learning through listening has always bulked fairly large in school, and the expansion of radio, television and sound movies as instructional media in recent years has increased the importance of the aural channel. Skills of listening are probably less directly related to school instruction than other communication skills, but appraisal of ability to comprehend and retain what has been heard is certainly of some educational importance.

#### MEASUREMENT OF STUDY SKILLS

Knowledge about and skills in obtaining information have found a place in most of the recent achievement tests. The emphasis in many

schools upon individual and group projects, and upon gathering information from all parts of a book—not just the words in it—has supported the need for tests dealing with such skills. In one way or another, most current batteries appraise such skills as: reading graphs and charts; reading tables; reading maps; picking appropriate reference sources; finding information in a reference source; using a dictionary. These skills are occasionally incorporated in tests of content areas such as social studies, science and mathematics (*STEP*) or included only as a brief subtest in some other major test (*California*), but usually some combination of them now appears as a distinct test in the battery, yielding one or more distinct scores. Thus the *Iowa* provides three subtests which between them cover all of the six skills referred to above. This is one major respect in which the widely distributed batteries of the present differ from those of the 30's—a place in which achievement tests have adapted to, and perhaps even helped to foster certain types of curriculum emphases.

#### MEASUREMENT IN CONTENT AREAS

The six batteries that we use as illustrations split evenly on the matter of including tests of information in content areas. Three (*STEP*, *Metropolitan*, *Stanford*) include tests in social studies and science; the other three do not. However, the three that do provide these content subtests also provide a “partial” battery, a more limited set of tests from which the content subtests have been omitted, in recognition of the fact that some schools are not interested in them.

Tests in content areas have tended to play a less prominent part in standardized testing in recent years, reflecting a feeling that in the common core of the curriculum items of information are a good deal less universal than skills. Thus, at one time a “Literature” section appeared in some tests, but it now has disappeared because of the feeling that the stories and books read will vary too much from school to school to provide any dependable common core in terms of which the schools may be compared. The common core is undoubtedly larger for science and social studies, but one may still question whether it is large enough to permit meaningful comparisons between different school systems. This has been handled in the *STEP* by making the science and social studies tests into reading and study skills tests in the science and social studies areas. That is, the role of information and specific knowledge in the area is held to a minimum, and the tests become primarily tests of ability to comprehend and work with ideas in the specific subject matter field.

## BATTERIES FOR HIGH-SCHOOL ACHIEVEMENT

The batteries that have been discussed so far are for the elementary school and junior high school. It is at these levels that survey batteries have been most widely used. The more departmentalized and specialized program of the high school and college appears to call more for specific tests in particular subject areas. The Cooperative Test Division of the Educational Testing Service markets a range of such specific achievement tests all tied to a common score scale.

There are, however, several comprehensive batteries at the secondary and higher levels. Summary information on four of these is presented in Appendix III. The four batteries suitable for secondary-school use have in common tests of content knowledge in natural sciences, social studies, and mathematics. Three of them also provide an evaluation of achievement in English, tending to emphasize correctness and effectiveness of expression. The *Sequential Tests of Educational Progress (STEP)*, which include a battery for grades four through six, as well as batteries for seven through nine, ten through twelve, and the first 2 years of college, emphasize communication skills, with objective tests of reading, writing, and listening, as well as a subjectively graded essay. The *Iowa Tests of Educational Development* go beyond the fields of content knowledge and undertake to appraise abilities to locate, read, and understand materials in the different subject areas, thus attempting to test ability to get and use knowledge as well as the amount of knowledge already obtained. Tests of this sort were found especially useful in evaluating the educational level of individuals much of whose education had occurred outside the usual school setting, specifically soldiers in World War II who had acquired various amounts and types of training while in military service. Evidence is reported by the test authors that score on this battery predicts college achievement at least as well as grades during 4 years of high school.

## USING THE RESULTS OF A SURVEY BATTERY

Since the survey achievement battery is one of the two or three most widely used types of standardized test, it is fitting that we consider ways in which the results from this testing may be used and appraise the soundness of each. Various things are done with the results from achievement testing, some useful, some relatively futile, and some perhaps positively harmful. Let us examine some of the possibilities. One possibility, of course, is that the tests are just given, scored, incorporated in some type of summarizing report, and filed away. This



is one of the forms of futility referred to above. We shall dismiss this possibility and assume that at least *something* will be done with the test results. Let us examine various uses that might be made of them.

#### USE TO EVALUATE THE CURRICULUM OF SCHOOL OR SCHOOL SYSTEM

As part of a total appraisal of the effectiveness of its program, a school system may well wish to include measures of progress in basic skills. An achievement battery provides a convenient tool for doing this. The results will show how well the particular school or school system has progressed on the several components of the battery in relation to the norming groups. However, in interpreting this progress, three cautions must be borne in mind.

1. The evaluation is only partial, not complete. The battery can give information only on the range of skills that it covers, and these skills represent only a fraction of the objectives of the modern school. Because they are so conveniently measurable, they may become overvalued. This is an insidious danger. The school system must seek to supplement standardized achievement tests with broader and more informal appraisals of other objectives if it is to obtain a well-rounded evaluation of its program.

2. Local emphases may differ from those that characterized the national sample. The particular school system may have placed heavier emphasis upon reading or may have delayed the introduction of formal instruction in arithmetic. In so far as local emphasis and effort are atypical, local accomplishment may be expected to be atypical. Evaluation of achievement in the single school or system must take account of distinctive local emphases.

3. Evaluation of pupil performance in a school must take account of the characteristics of the pupil population. Schools, communities, even regions differ in the economic and cultural level of the population served. Associated with these differences are differences in average level of ability as measured by our intelligence tests. The expectancy for achievement must be tempered to take these factors into account. This may be approximated by developing regional norms or norms for schools of a particular type.

#### USE TO PLAN THE PROGRAM FOR A CLASS GROUP AND THE PUPILS IN IT

Every fall each teacher in most schools faces a new group of pupils. Within the limits set by the course of study (which may in some instances be quite rigid limits) he must plan a program of activities for the group as a whole and must adapt that program as best he can to

each of the children in the group. He must decide where to pick up the various skill subjects, how much time to devote to review of materials presumably taught in the previous year, and how fast to move ahead. He must plan appropriate enrichment experiences and materials for independent work and free time. He will probably want to form informal groupings within the class for work together at a common level.

To do these things he needs to get to know the pupils in the group as quickly, thoroughly, and accurately as possible. Administration of a standard achievement battery is an efficient way of laying the foundations for that picture of the class which will permit him to adapt his plans to the individuals with whom he has to deal. The scores will provide a guide as to whether the group as a whole is superior, average, or retarded in each of the basic skills he is trying to develop. They may indicate group areas of relative strength and weakness. They will pick out the children who could profit from more challenging tasks than those presented to the class as a whole, those who need less demanding materials, and those who should be considered for special help either within the classroom or through a remedial teacher if one is available.

It should be understood that this function of informing the teacher about his pupils is not to depend on tests alone. Every contact with the children helps the teacher to get a "feel" for the class group and the pupils in it. A richness of understanding of individual pupils can only come from working with them as persons. But the set of standard test scores provides an objective reference framework within which to see the rest of the picture of the class and the pupils. This function can, of course, be served by tests given the preceding spring and forwarded to the teacher when he meets the class in the fall. Technically, test results from spring testing would be quite serviceable, since pupils' skills are not likely to shift around greatly during the few summer months. But it is likely that tests given early in the fall will seem more current and alive than results from the preceding spring and will be more likely to be used by the teacher in determining his plans for the class.

#### USE TO IDENTIFY INDIVIDUALS FOR MORE DETAILED STUDY

One function of an achievement battery is to help screen out a fraction of the group of children for more intensive study. Though every child should be studied as an individual, there are in every school system some children more in need of special help than others. In those cases in which the symptom is failure to progress in school skills, the

problem may be first identified by poor performance on a standardized test.

Gross irregularities in performance on different subtests, performance far below his age or grade level, or performance well below his aptitude as indicated by an intelligence test are cues suggesting further study. But they are only cues. They are only symptoms suggesting that something may be wrong. The significance of the symptom must be investigated further. In the first place, the educational achievement must be related to a measure of aptitude to see that the child is falling behind what should be expected of him. Where it is reading achievement that is at issue, his achievement should be related to performance on an aptitude test not involving reading. Then if the deficiency appears to be a specific retardation in some school skill, further diagnostic procedures need to be applied to determine the exact nature and causes of the deficiency.

#### UNDERSTANDING THE INDIVIDUAL PUPIL

Though special study and remedial activities may be possible for only part of the children in a class, the school and teacher have the responsibility of knowing every child as well as possible so as to provide the best possible guidance for him in his present school activities and in plans for the future. Level of educational achievement is one facet of the picture that is needed in understanding and guiding each pupil. Appraisal of present adjustment, planning for future education, and counseling about a life career can all be helped by information about educational progress.

#### MAKING UP CLASS GROUPS AND PLACING INDIVIDUAL PUPILS

In a large school where there are enough pupils to fill several classes in a grade or several sections in a subject, some procedure must be adopted for assigning pupils to particular groups. Fashions with respect to grouping together children of similar ability have changed several times over the past 50 years. At present, this procedure seems to be a relatively respectable one in educational circles. When the basic decision has been made to try to achieve homogeneous groups within each classroom, a standardized achievement test or battery provides one useful tool for achieving this end.

Of course the term "homogeneous group" is rather misleading, because the most we can do is to make a group somewhat less heterogeneous. Whether we use over-all level on an achievement battery, reading level, score on a scholastic aptitude test, or some combination of these, the children in any group will still vary markedly. They will

vary in part because it will always be necessary to include children with a range of scores in any group. They will vary in even larger part because different abilities are not perfectly correlated. The child who is most outstanding in reading may be fairly mediocre in arithmetic or spelling, and vice versa. Grouping will not do away with the need to treat pupils as individuals, or to group them *within* the class for some special purposes, but it may reduce the range of individual differences enough so that the whole group can work together better and participate effectively in common academic enterprises. At the high-school level, where separate grouping is possible by subject areas, a specific measure of achievement in the subject area is likely to provide a more useful basis for grouping than a measure of over-all achievement.

The problem of placement in a class also arises for transfers into a school system. Here it may be a question not only of the section into which to place the pupil, but even of the grade level at which he can perform adequately. Results from standardized achievement tests can help in this decision. They make it possible to compare his achievement with that of the groups into which he may be placed in a way that is not possible from school marks alone.

#### EVALUATING THE TEACHER

It is reported that in some school systems a standardized achievement battery is used, either openly or covertly, to evaluate the success of the teacher. He is judged by the performance his class shows on standardized tests given at the end of the school year. He is expected, with varying degrees of unrealism, to bring his class "up to the norm" on these tests.

This procedure seems questionable at best, and quite possibly vicious. It fails to take account of a number of important considerations. In the first place, the achievement of a class group is a function of their whole previous educational history, not merely of the year just past. It is unreasonable to hold the teacher who has taught a group for a single year solely responsible for their present status. In the second place, achievement depends upon aptitude and upon out-of-school cultural experiences as well as upon schooling. Unless the evaluator is prepared to make an appropriate adjustment for the intellectual and socio-economic level of a particular class—and class groups can differ widely in these respects—no reasonable base-line can be provided for evaluating what the teacher has accomplished. In the third place, the skills measured by an achievement battery represent only a fraction of the objectives of a modern school. Comparison of teachers

with respect to this partial criterion neglects much of their work and may provide a very unfair evaluation of relative worth of two teachers whose strengths lie in different directions. Fourth, placing a premium upon easily testable skills when evaluating the teacher is almost inevitably going to lead the teacher to overvalue those skills in his teaching. As he is judged, so will he judge. Skills will tend to become the one central theme of his teaching, at the expense of all the other outcomes the school is trying to achieve. He will, with varying degrees of directness, teach for the tests. Finally, one may mention the demoralizing effect upon teachers of a mechanical, external evaluation that is subject to all the technical limitations discussed above.

### SUMMARY STATEMENT

The typical standardized achievement test is superficially much like an objective test made by the classroom teacher. However, it is based on large segments of knowledge or skill common to the programs of many schools, and it provides norms. These features mean that it is appropriately used in making broad comparisons—between schools or classes, between areas of achievement, or between achievement and aptitude.

Just as an analysis of the objectives to be measured was indicated as the first step in thoughtful construction of a classroom test, so an analysis of objectives is a prerequisite for evaluating a published test. The test can only be evaluated in terms of the objectives that the teacher or school is trying to achieve.

Most widely used standardized tests are survey tests, giving a general appraisal of level of accomplishment in a broad area. If the teacher is to work constructively with the pupil, such survey results need to be supplemented by more specific and diagnostic information. Some published diagnostic tests exist, and these can be supplemented by informal teacher appraisals. However, the reliability of difference scores and consequently of differential diagnoses is often low. Diagnostic clues should be considered quite tentative.

Certain skills, such as those of handwriting, shop work, or domestic arts, can be appraised effectively by comparing a pupil product with a scaled set of standard samples.

Standardized achievement test batteries are very popular for school use. In these the advantage of unity in plan and standardization must be weighed against the inflexibility of a single total battery. The published batteries are similar in general design, though they differ in (1) content subjects included, (2) emphasis on work-study skills, (3) bal-

ance of emphasis among different areas, and (4) specific pattern of items in each field.

When used with discretion and proper reservations, a standardized achievement battery can serve a useful purpose as *one* type of evidence (1) to evaluate a school's educational program and its several components, (2) to help the teacher plan the work of his class and the grouping of pupils within it, and (3) to provide an understanding of the individual pupil. Standardized test results should rarely, if ever, be used as a basis for evaluating the effectiveness of individual teachers.

### SUGGESTED ADDITIONAL READING

- Harris, Chester W., Editor, *Encyclopedia of educational research*, 3rd ed., New York, Macmillan, 1960, pp. 72-73, 464-466, 804, 881-883.
- Katz, Martin R., *Selecting an achievement test: principles and procedures*, Princeton, Educational Testing Service, 1958.
- Lindquist, E. F., and A. N. Hieronymus, *Manual for administrators, supervisors, and counselors, Iowa Tests of Basic Skills*, Boston, Houghton Mifflin, 1956.
- Sequential tests of educational progress, teacher's guide*, Cooperative Test Division, Educational Testing Service, Princeton, N. J., 1958.
- Traxler, Arthur E., *The use of test results in diagnosis and instruction in the tool subjects*, rev. ed., Educational Records Bulletin No. 18, New York, Educational Records Bureau, 1949.
- Traxler, Arthur E., et al., *Introduction to testing and the use of test results in public schools*, New York, Harper, 1953, pp. 89-95.

### QUESTIONS FOR DISCUSSION

- For which of the following purposes would a standardized test be useful? For which should a teacher expect to make his own test? Why?
  - To determine which pupils have mastered the addition and subtraction of fractions.
  - To determine which pupils in a class are below standard in arithmetic computation.
  - To determine the subjects in which each pupil in a class is strongest and weakest.
  - To determine for a class which punctuation and capitalization skills need further teaching.
  - To form subgroups in a class for the teaching of reading.
- Examine some standardized reading test. In view of the tasks it presents, which of the objectives outlined on pp. 291-292 does it measure adequately? Which does it measure to some extent? Which does it fail to measure at all?
- Examine a standardized achievement test for a subject that you are

teaching or plan to teach. Which of the objectives that are important in the subject are measured adequately by the test? Which ones are not?

4. Make a critical comparison of two achievement test batteries for the same grade. How do they differ? What are the advantages of each from your point of view?

5. What are the advantages and disadvantages of a dictation as opposed to a multiple-choice type of spelling test?

6. Suppose you are teaching mathematics in the first year of junior high school. List the steps you would take to diagnose the achievement level of the pupils and plan for remedial instruction.

7. The manual of test W states that it can be used for diagnostic purposes. What should you look for to determine whether it has any real value as a diagnostic aid?

8. Why should we be specially concerned about the reliability of the scores resulting from a set of diagnostic tests? What implications does this have for using and interpreting such tests?

9. Suppose that you are a college chemistry teacher and are interested in the laboratory skills of glass blowing that your students have developed. How might you develop a product scale for evaluating their skill?

10. Before you can make a sound evaluation of the grade equivalents made on a battery of achievement tests by a class or pupil, what information do you need beside the converted scores themselves?

11. The town of M gives the *Stanford Achievement Tests* to pupils in grades four and six and records on the cumulative record card only the grade equivalent for the whole test. What are the disadvantages of this type of record?

12. You have given a standardized achievement battery in October to your fourth-grade class. What might you, as teacher, do on the basis of the results?

13. In city K, the *Metropolitan Achievement Test* is given to all schools in April. The average grade level for each class group and for each subject is reported to the superintendent of schools' office, and these results are mimeographed and distributed to all schools. What are the gains from this procedure, and what are the dangers in it? What changes would you suggest?

14. In a fourth-grade group you have data from a group intelligence test and from an achievement test battery. On what basis would you select individuals to receive special remedial work, either in your class or with a special teacher? What are the hazards of this procedure?

15. What should be the role of standardized test results in evaluating the performance of the classroom teacher?

## Chapter 12



# Questionnaires and Inventories for Self-Appraisal

The last three chapters have been devoted to measures of ability: what the individual *can* do under test conditions and motivation to do his best. We shall move on now to measurement of other aspects of personality—to the appraisal of what he *will* do under the natural circumstances of life. Both in our discussions of personality and in our efforts to develop instruments of appraisal, we must recognize that the person is a unified whole. Any aspects or traits that we may separate out are separated out for our convenience. They do not exist as separate entities. They are only aspects of or ways of looking at the unitary person. However, it is inevitable that we do pick the person to pieces to study and understand him. We cannot look at everything at once.

In Chapter 2 we identified five segments of personality; to wit:

*Temperament* refers to the individual's characteristic mood, activity level, excitability, and focus of concern. It includes such dimensions as cheerful-gloomy, energetic-lethargic, excited-calm, introverted-extroverted, and dominant-submissive.

*Character* relates to those traits to which definite social value is attached. They are the "Boy Scout" traits of honesty, kindliness, co-operation, industry, and such.

*Adjustment* is a term that we shall use to indicate how well the individual has been able to make peace with himself and the world about him. In so far as the individual can comfortably accept himself and his world, in so far as his ways of life do not get him into trouble in his social group, he will be considered well adjusted.

*Interests* refer to tendencies to seek out and participate in certain activities.

*Attitudes* relate to tendencies to accept or reject particular groups of individuals, sets of ideas, or social institutions.



## METHODS OF STUDYING PERSONALITY

Most of the evaluation techniques we shall consider in this and the following chapters have to do with one or more of the aspects of personality identified above. To what sources may we go for evidence on these aspects when we wish to study an individual? First, we can see what the individual has to say about himself. Second, we can find out what others say about him. Third, we can see what he actually does, how he behaves in the real world of things or people. Fourth, we can observe how he reacts to the world of fantasy and make-believe.

### WHAT THE INDIVIDUAL SAYS ABOUT HIMSELF

One obvious source for information about a person is that person himself. No one else has as intimate and continuous a view of Johnny as Johnny has of himself. He is aware of hopes and aspirations, worries and concerns that may be well hidden from the outsider. To get at the individual's view of himself we may interview him, probing those areas that seem sensitive or significant. Another approach is to incorporate the questions that might be asked in a face-to-face interview into a uniform questionnaire or personality inventory. The choices the individual makes in responding to the set of questions are scored in various ways to provide a picture of him as he describes himself. These procedures will be elaborated in this chapter, and their strengths and weaknesses pointed out.

### APPRAISAL THROUGH THE OPINION OF OTHERS

For some purposes, we may be interested in how a person is perceived by his fellow beings. Is he seen as a friendly fellow worker? A fair teacher? An industrious pupil? A convincing salesman? A generally desirable employee? The opinion of others may be the significant fact in certain settings. It is also a very convenient way of getting a summary appraisal of a fellow man. For these reasons, rating procedures have been widely used. We shall consider their values and limitations in the next chapter.

### MEASURES OF BEHAVIOR

It can be argued that for practical purposes an individual's personality is what he does, rather than what he says or what is said about him. The problem is to develop procedures for appraising genuine behavior, not distorted for the purpose of making a good impression. Some attempts have been made to do this with objective tests, and we

shall consider these briefly in Chapter 14. Of more importance and current interest are procedures for observing the individual and for recording or evaluating his responses as they are seen by an observer.

#### THE WORLD OF IMAGINATION AND FANTASY

What an individual will tell about himself in response to questions is limited by his willingness to reveal himself, his understanding of himself, and his understanding of the language in which the questions are presented. For this reason, indirect methods have been sought to avoid these limitations and permit him to "open up" more fully. One indirect avenue is that of fantasy, imagination, and make-believe. We may study what the person sees in ink blots, what stories he tells about an ambiguous picture, what play scenes he acts out with dolls, what he does with paints and modeling clay. These materials and others have been used to elicit imaginative productions that psychologists have studied as a source of understanding of children and adults. The individual is allowed to express himself through play materials or to project his own interpretations into ambiguous stimuli, and thus to reveal himself to us. These are expressive and projective techniques for personality appraisal. We shall undertake to describe and evaluate them in Chapter 15.

#### INTERVIEW

If we wish to find out about a person, one obvious way to do so is to ask him questions and evaluate his answers. If the questions are asked orally in a face-to-face situation, we are carrying out an interview. The interview has been a perennial favorite as a way of studying people. It is widely used by colleges and professional schools, by employers, and by clinicians working with disturbed individuals. Why is the interview looked upon with such widespread favor?

The popularity of the interview is *not* based primarily upon its demonstrated validity as a device for appraising people. In fact, evidence for the validity of the impressions or conclusions derived from interviews is spotty and rather contradictory. Interview procedures are basically subjective, variable and heavily dependent upon the skill of the interviewer. It has repeatedly been demonstrated that different interviewers interviewing the same person come up with quite varied impressions of him. The variability arises in part from variation in the questions asked and the lines of inquiry intensively pursued. It arises in part from differences in interpretation and evaluation of the

responses the individual makes. The typical interview is not a precise or efficient psychometric technique.

The appeal of the interview lies rather in its great flexibility and adaptability. The interviewer can structure the interview in whatever way seems to him most suitable, in the light of the purposes of the interviewing and of the responses elicited to prior questions. He can skim over certain areas; probe intensively in others. He can give full play to his "clinical insight" and "intuition."

There is no doubt that the flexibility possible in the interview situation has certain elements of strength. It permits the wise interviewer to take full advantage of everything he has learned about the interviewee as he directs the further course of the interaction. But this same flexibility contains elements of weakness. It tends to destroy comparability from one interviewer to another and from one interviewee to the next. It makes it possible for an interviewer to ride a personal hobby and ignore many obvious areas of inquiry. Just as it permits full scope to the wisdom of the wise, so also it gives abundant rope to the foolishness of the foolish or the biases of the biased.

One approach that undertakes to reduce the subjectivity and variability of interview procedures, while still maintaining the flexibility and vividness of direct personal contact, is the *structured interview*.<sup>4</sup> Structured interview procedures give the interviewer a fairly detailed guide of topics to be covered and areas of inquiry to be included. These may include, for an employment interview, family patterns and interrelationships, school interests and activities, sports participation, previous work history, reasons for leaving previous jobs, and other similar areas. Within any one of these rather broad areas there may be several more specific questions to which the interviewer is to find an answer. The interviewer retains freedom and flexibility with respect to the order in which he attacks the different topics and the depth to which he pursues each. At the same time, he has a guide to make sure that a standard set of areas of inquiry is covered in each interview. The structured interview is a compromise between the free interview on the one hand and the printed biographical data blank or personality inventory on the other.

The questions during an interview session elicit responses that are descriptive of the individual. However, these responses require some degree of interpretation if they are to provide a useful picture of him. The interpretation may sometimes flow quite directly from the manifest content of the responses. This would be the case when the interviewer interprets a report of membership in many school organizations, the holding of many school offices, out-of-school experience in selling, and membership in the debating team as evidence of assertiveness and

social leadership. Sometimes the interpretation may be quite indirect, and dependent upon the latent or concealed, rather than the manifest or obvious content of the response. This is true of many of the psychoanalytic interpretations of the communications from patient to analyst. In these variable and unstandardized interpretations lie potential strengths and frequent weaknesses of the interview as an appraisal technique.

The clinical interview is an unstandardized inquiry, highly dependent upon the particular interviewer both for the way it is carried out and for the way it is interpreted. Furthermore, individual interviews place very heavy demands upon the time of interviewing personnel, demands which may be prohibitive in a number of situations. To economize on interviewer time, then, and to provide an inquiry that is uniform in presentation and procedure for evaluation, the printed questionnaire has been developed. The self-report questionnaire or inventory is essentially this: a standard set of questions about some aspect or aspects of the individual's life history, feelings, preferences, or actions, presented in a standard way and scored with a standard scoring key.

### THE BIOGRAPHICAL DATA BLANK

An obvious and important use of the questionnaire is as a means of eliciting factual information about the individual's past history. Place and date of birth, amount and type of education and degree of success with it, nature and duration of previous jobs, hobbies, special skills, and a host of other biographical facts can be determined most economically through a blank filled out by the individual himself. It is the economy and efficiency of this approach that makes it particularly appealing. Though his reports may be inaccurate in some respects, the individual himself is probably the richest single repository for the factual information we would like to have about him.

The problems in using questionnaires to elicit facts are primarily problems of communication. When questions are preformulated and appear in printed form and answers are written down, misunderstanding may occur either in the respondent's interpretation of the question or in the using agency's interpretation of his response. If there is no personal interaction, these misunderstandings cannot be cleared up with an oral question or a further probing into the area of uncertainty. It is important, therefore, that a fact-finding questionnaire be very carefully worded and that it be tried out in preliminary form with small groups to make sure that the ambiguities have been cleared out of it.

An interview to supplement the questionnaire is often desirable in order to permit clarification of any of the responses to questionnaire items that are puzzling to the user or to get fuller information on some points. As a matter of fact, one appropriate use of self-report inventories of all types is to provide a jumping-off place for an interview, the questionnaire providing leads that may be followed up in the interview.

Sometimes the factual information on an application blank or other fact-finding questionnaire has been used to determine whether the individual meets certain stated requirements to be eligible for a job, educational program, or the like. Sometimes it has been used as part of the raw material from which the personnel officer, director of admissions, or scholarship committee makes a clinical judgment of the individual's desirability as an employee or student. In a few instances, however, biographical data blanks have been analyzed item by item to determine to what extent particular responses to each item actually predict some criterion of job success. Items found to discriminate more successful from less successful individuals are given a score credit, and the separate items are summed to give a score for the blank as a whole. Thus, the World War II programs for selecting pilot trainees for both the Army and the Navy used a scored biographical data blank that was treated just as if it were a test. The life-insurance companies have for a number of years used an *Aptitude Index* in selecting insurance salesmen, one section of which consists of factual items about the individual applicant. Thus, the individual is asked about the amount of insurance he himself carries, his net worth, etc. A scoring system assigns scores for each response in terms of the success experienced by those in the validation group who had given that response.

In the examples given above, objective scoring of a biographical data blank provided one of the most valid predictors of job success. These results suggest that there may be a number of other selection situations in which a standard scoring procedure could be used with advantage. The development of scoring weights is a major undertaking, but once a scoring system has been developed the scoring of individual blanks proceeds rapidly. It has even been possible in military use of biographical inventories to prepare them in multiple-choice form and score them like any standard test.

### INTEREST INVENTORIES

One aspect of the individual's make-up that we would like to study, both to understand him as a person and to help in such immediately

practical problems as educational and vocational guidance, is the domain of interests and aversions, preferences for activities and surroundings. Of course, in the matter of vocational interests, the simplest procedure would seem to be to ask the individual how much he would like to be an engineer, for example. However, this doesn't work out very well in practice. In the first place, people differ in the readiness with which they exhibit enthusiasm. "Like very much" for person A may signify no more enthusiasm than "like" for person B. In the second place, people differ substantially in the nature and completeness of their understanding of what a particular job means in terms of activities and conditions of work. "Engineer" to one person may signify primarily out-of-doors work; to another it may carry a flavor of the laboratory or drafting board; to still another it may signify vaguely a high-prestige, science-oriented job. These varied and incomplete meanings cause a response to the single question, "How much would you like to be an engineer?" to be a rather unsatisfactory indicator of the degree to which the individual has interests really suitable for the profession of engineering. It is for these reasons that psychometricians have undertaken to broaden the base of information and to ask a whole array of questions about the individual's likes and dislikes, rather than simply to ask directly about preference for particular jobs.

#### THE STRONG VOCATIONAL INTEREST BLANK

One of the best known instruments for appraising interests is the *Strong Vocational Interest Blank for Men*. This inventory is made up of 400 items, broken up into the following types: liking for occupations, liking for amusements, liking for activities, reaction to peculiarities of people, choice or preference between activities, and evaluation of personal abilities and characteristics. To most of the 400 items in the *Strong Blank* the individual responds by marking one of the three given options L, I, and D (Like, Indifferent, Dislike). A response is called for to each item. Over 40 different scoring keys have been developed for the men's blank. Most of these are for specific occupations, largely at the professional level, such as architect, chemist, lawyer, or YMCA secretary, though there are also keys for interest maturity, masculinity of interests, and occupational level.

The scoring key for each occupation was developed by comparison of a group of men who were successfully engaged in that occupation with a reference group of men-in-general. Thus, the per cent of men in occupation A choosing the L, I, and D options to item 1 is compared with the per cent of men-in-general choosing these same options. If enough more men in occupation A choose a particular option, that

option receives a plus score for occupation A. If the per cent is smaller for occupation A, the option receives a minus score. If the per cent for occupation A is very much larger or smaller than for men-in-general, the score may be as much as +4 or -4. Smaller scores are assigned to smaller differences. Thus, responses are weighted to take account of the sharpness with which the item discriminates.

Table 12.1 shows the scoring key for the first ten items in the blank for four different occupational keys. Note the range of weights for the different items. Note that some or all of the options for a given item may receive a zero weight.

Table 12.1. Scoring Weights for Sample Items and Keys of Strong Vocational Interest Blank for Men

Item	Scoring Key											
	Engineer			Social Science Teacher			Farmer			Production Manager		
	L	I	D	L	I	O	L	I	D	L	I	D
Actor (not movie)	-1	0	1	1	0	-1	0	0	1	0	0	0
Advertiser	-2	0	2	0	1	-1	-2	1	1	-1	0	1
Architect	2	-1	-1	-1	0	1	0	0	0	0	0	0
Army officer	1	0	-1	1	0	-1	0	0	0	1	0	0
Artist	0	0	0	-1	0	0	-1	0	1	-1	0	0
Astronomer	1	0	-1	-1	0	0	-1	0	1	0	0	0
Athletic director	-1	1	0	2	-1	-2	0	0	0	0	0	-1
Auctioneer	-1	-1	2	0	1	-1	0	1	-1	0	0	1
Author of novel	-1	1	0	1	0	-1	-1	0	1	-1	0	0
Author of technical book	3	-1	-2	0	1	0	-1	0	1	1	0	-1

An individual's score is obtained by summing up the plus and minus credits corresponding to the responses he has chosen. Since the weights are different for each occupation, a separate scoring key is required, and a separate score is obtained for the examinee for each scale. Thus, a series of scores is obtained showing how closely the responses given by our examinee correspond to those typically given by each specific occupational group. Raw scores are translated into a standard score scale in which 50 represents the mean for men in the specific occupation. A scale of letter grades is also provided, in which A represents close resemblance to the particular occupational group, B+, B, and B- lesser degrees of resemblance, and C+ or C interest patterns quite different from those of the particular occupational group. Table 12.2 shows the standard scores and letter ratings on the occupational scales of the blank for one college freshman. This young man shows interest patterns resembling closely (A) those of chemists,

Table 12.2. Scores on *Strong Vocational Interest Blank* for a College Freshman

Occupation	Standard Score	Letter Rating
I. Artist	26	C+
Psychologist	22	C
Architect	29	C+
Physician	42	B+
Dentist	41	B+
II. Mathematician	26	C+
Engineer	44	B+
Chemist	52	A
III. Production manager	39	B
IV. Farmer	59	A
Carpenter	44	B+
Math and science teacher	48	A
V. YMCA physical director	34	B-
Personnel manager	21	C
YMCA secretary	Low *	C-
Social-science teacher	17	C
City school superintendent	Low *	C-
Minister	Low *	C-
VI. Musician	25	C+
VII. CPA	16	C
VIII. Accountant	25	C+
Office worker	25	C+
Purchasing agent	28	C+
Banker	22	C
IX. Sales manager	19	C
Real estate salesman	17	C
Life-insurance salesman	Low *	C-
X. Advertising man	19	C
Lawyer	20	C
Author-journalist	24	C

\* "Low" designates a standard score of 15 or lower.



farmers, and mathematics and science teachers. His interests are also quite like (B+) those of physicians, dentists, engineers, and carpenters. His interests are very *unlike* (C-) those of YMCA secretaries, city school superintendents, ministers, and life-insurance salesmen.

Strong has developed a companion *Vocational Interest Blank for Women* that follows closely the pattern of the blank for men. However, the blank has been rather less thoroughly developed than the men's blank, and seems to have been rather less successful. This may be due to the fact that specifically vocational interests are less central in the lives of many women, being contaminated by general "homemaker" interests, so that interest profiles in women tend to be less clear-cut and meaningful.

Originally, scoring the *Strong Vocational Interest Blank* was a very time-consuming task because of the large number of different scores that are called for. Hand-scoring a blank was a matter of several hours' work. But twentieth century electronics has hit the test-scoring field, and a special device developed by E. J. Hanks has made it possible to score the blanks at very high speed. This scoring machine is available only at Engineers Northwest, Minneapolis, Minnesota. The special answer sheets must be sent to this organization, where they will be scored at a cost that is a fraction of what the cost would be by hand methods.\*

There are two points about the construction of the *Strong Blank* to which we wish to call especial attention at this time. In the first place, the person taking the test responds by choosing one of a set of response categories for each item (L, I, D). A particularly effusive individual *could* choose all L's, and a particularly jaundiced one *could* choose all D's. There is a certain amount of freedom to impose one's own standards upon the task. Secondly, the keys are externally determined. That is, they are defined by the responses of a particular job group and not by any internal logic. We wish now to contrast with the *Strong Blank* the *Kuder Preference Record*, which is different with respect to both of these features.

#### THE KUDER PREFERENCE RECORD (VOCATIONAL)

The *Kuder Preference Record (Vocational)* is made up of triads, or sets of three options. Typical sets might read:

- Go for a long bike in the woods.
- Go to a symphony concert.
- Go to an exhibit of new inventions.

\* A price of 70¢ per answer sheet was quoted in 1960 for scoring blanks in quantity lots.

Fix a broken clock.  
 Keep a set of accounts.  
 Paint a picture.

In each set the individual is required to mark the one he would like to do *most* and the one he would like to do *least*.

Scoring keys were established on the basis of the *internal* relationships of the items. Thus, a study of the responses to the items showed that a number of items dealing with mechanical activities tended to hang together. If a person chose one he was likely to choose others, and if he rejected one he was likely to reject the others. Moreover, items in this group showed relatively little relationship to the remaining items. The items grouped together in a distinct cluster. From the nature of the items it was evident that this cluster related to mechanical interest. Those items having a substantial correlation with this cluster were included in a scoring key that gave a score for mechanical interest.

In the same way, other clusters were identified and built up in which the items went together but were largely independent of items not in the cluster. Scoring keys were developed for these. The *Preference Record* now yields scores for the following interest clusters: outdoor, mechanical, computational, scientific, persuasive, artistic, literary, musical, social service, and clerical. Raw scores are converted into percentiles, separate norms being supplied for male and female high-school students and for male and female adults.

In Table 12.3, the *Kuder* scores are given for the same college freshman whose *Strong* scores were shown in Table 12.2. On the *Kuder*,

Table 12.3. *Kuder Preference Record Scores of a College Freshman \**

Interest Area	Raw Score	Percentile Equivalent
Outdoor	71	95
Mechanical	58	87
Computational	17	16
Scientific	60	93
Persuasive	25	07
Artistic	30	68
Literary	23	78
Musical	12	45
Social Service	36	46
Clerical	19	01

\* Scores for same individual shown in Table 12.2.

this young man stands highest on outdoor, scientific, and mechanical interest. He is very low on clerical and persuasive interests. These findings can be studied in relation to his interest in specific occupations, as shown in Table 12.2. The two sets of results are obviously consistent and support one another.

#### COMPARISON OF STRONG AND KUDER INVENTORIES

Note that in the *Kuder Preference Record*, the examinee is forced to pick a most liked and a least liked activity in each set. No matter how much or how little he likes all three, one must be preferred and one rejected. This forced-choice pattern appears in a number of inventories and should be contrasted with the category-response pattern found in the *Strong*. The forced-choice pattern forces a common frame of reference upon everyone. Differences in general optimism are controlled. Everyone must express the same number of preferences and rejections. Thus, superficial differences in standards of judgment, or what has been called "response set," are eliminated. But so also are genuine differences in interest level. Whether the forced-choice pattern produces a net gain in this respect is still a matter of debate.

Note again that in the *Preference Record* the several scores relate to coherent interest clusters rather than to something outside the individual or the test. The scores carry their own relatively direct meaning in terms of the common theme running through the cluster of items. The meaning does not have to be inferred by thinking what lawyers or salesmen are like. If our purpose is to build up a meaningful description of an individual, the internally consistent scales appear more satisfactory than those that are externally oriented. To say that a person is high on mechanical, scientific, and out-of-doors interests and low on clerical and persuasive is more directly interpretable than to say he is high on interests characteristic of farmers, chemists, and mathematics-science teachers and low on those characterizing ministers and YMCA secretaries. Internally coherent clusters definable in terms of their common theme "make sense" better than job-oriented appraisals.

When it comes to rating the individual for a specific job, however, the balance of advantages is radically changed. If our concern is to help the individual decide whether he would be content in the job of engineer, it is much more directly relevant to know how well his interests correspond to those of successful engineers than to know how high his mechanical and scientific interests are. In the first case, the scoring key itself defines what the interests of engineers are; in the second case we must either infer this or determine it from a separate study.

Either the internally consistent or job-oriented approach to inventorying interests is possible; which will work better depends on our particular purpose. If our purpose is to appraise appropriateness of interests for a limited number of specific jobs, this may be done effectively with a specific job key for each job. If, however, our concern is to get a meaningful description of a person and perhaps to be prepared to use that description to make inferences as to his suitability in any one of a very large number of jobs, then the homogeneous cluster scores seem preferable.

#### RELIABILITY, VALIDITY, AND PERMANENCE OF INVENTORIED INTERESTS

The *Strong Vocational Interest Blank* is one of the most thoroughly investigated psychometric tools we have, and, though the history of the *Kuder Preference Record* is shorter, it too has been intensively studied. Both instruments yield scores that are reasonably reliable for individuals in their teens or over. Thus, for 285 Stanford University seniors Strong<sup>16</sup> reports odd-even reliabilities for the separate occupational scales ranging from .73 to .94, with an average value of .88. A number of reliability studies with the *Kuder*, based on analysis of a single testing, give values averaging about .90. The reliability of the scores extracted from these interest inventories compares favorably with that of scores on ability tests.

For the *Strong*<sup>9,12,13,14,16</sup> there is evidence that interests show a good deal of stability over time, at least in adolescents and adults. Data on the average correlation at different ages and over different periods may be summarized as follows:

	<i>Upper Elementary School</i>	<i>High School</i>	<i>College Freshmen</i>	<i>College Seniors</i>
1 or 2 years	.55	.65	.80	
3 to 5 years	.30		.75	.75
6 to 10 years		.50	.55	.70

The stability is low in the elementary school, but for persons of college age stability compares favorably with that for intelligence tests.

In appraising the validity of an interest inventory as a description of how the individual feels about activities and events in the world about him, the main issue is the truthfulness of his responses. There isn't really any higher court of appeal for determining a person's likes and preferences than the individual's own statement.

A number of studies have indicated that inventories such as the *Strong* "can be faked. If a group of examinees is told to try to re-

spond the way that life-insurance salesmen would, they are generally rather successful in making themselves appear like life-insurance salesmen. However, this is no indication that the blank *will* be faked, even when used as an employment device.

When the inventory is used for counseling and to help the respondent, as is most often the case, there is probably little reason to anticipate intentional faking. The individual may be expected to report his likes and dislikes as he knows them. His self-knowledge is perhaps imperfect, so his reports may be inaccurate in some respects. Thus, he may say that he would like to attend symphony concerts because he feels that that is the thing to say, but his actions may belie his statement; he may in fact avoid concerts whenever they come his way. This lack of self-insight is a real problem. But it is probably mitigated somewhat, in the inventory approach to interests, where isolated points of poor insight will have only minor effects upon a final score.

The validity of interest inventories as predictors of later behavior is another matter. Scoring keys for the *Strong* were established by comparing men who were already in the occupation with men-in-general. *Kuder* occupational interest profiles have also been prepared by determining the average level in each of the interest areas for individuals already working in the occupation. But the common interest patterns of individuals in a field of work may have grown out of their work. The men may have come to exhibit certain common patterns from the very nature of their work experience. The crucial evidence on predictive validity would come from testing a group *before* they entered the world of work and determining whether those who later entered and continued in a particular occupation had distinctive interest patterns *before* they entered the occupation. This is an expensive operation, expensive in the time that must elapse before men can become settled in their occupation and expensive in the dissipation of cases among literally hundreds of occupations.

Strong<sup>15</sup> has been able to follow some groups who were tested as college undergraduates and does have some evidence on the extent to which students with interests characteristic of a particular occupation tended to enter that occupation and to persist in it. For the typical individual, the occupation in which he was actually working 10 years later ranked second or third for him among all the scales of the *Strong*. Considering group averages, those who remained in an occupation received higher interest scores for that occupation than for any other occupation and higher than those who switched to some other occupation.

McCully<sup>16</sup> followed up a group of men who had been given the *Kuder* as a part of Veterans Administration counseling at the end of

World War II. They were located several years later, and their occupation determined. Table 12.4 shows the average standard scores on each of the ten *Kuder* interest areas for those occupational groups that were large enough to justify study. The results show clear-cut

Table 12.4. Mean *Kuder* Standard Scores of Different Occupational Groups \*

	Mechanical	Computational	Scientific	Persuasive	Artistic	Literary	Musical	Social Service	Clerical
Accounting and related	-74	152	-32	37	-82	19	2	-14	118
Engineering and related	56	43	82	-16	7	1	-21	-46	-41
Managerial work	-28	41	-13	56	-27	19	-13	-3	42
Clerical— computing and recording	-27	87	-9	9	-50	4	3	-14	64
General clerical work	-19	-3	-31	-9	-14	22	3	17	30
Sales—higher	-65	-14	-40	111	-54	38	17	18	30
Sales—lower	-19	-12	-25	79	-32	10	6	13	16
General farming	32	-23	-16	-37	-4	-49	-42	12	-10
Mechanical repairing	81	-21	3	-40	28	-28	-30	-40	-29
Electrical repairing	60	-3	27	-35	5	-41	-13	-19	-29
Wrench crafts (fine)	63	-5	12	-24	38	-23	-20	-33	-2

\* Based on a mean of 0 and a standard deviation of 100 for the reference group of 2797 employed veterans.

and fairly substantial differences in pattern of interest for different occupations. Thus, evidence with respect to both the *Strong* and the *Kuder* indicates that they have a certain amount of validity as predictors of occupational choice.

#### INTEREST AND ABILITY

It is important not to confuse measures of interest and ability. The fact that a boy scores high on the scientific interest scale of the *Kuder* or on the physicist scale of the *Strong* is no guarantee that he possesses the intellectual and other aptitudes required to master the concepts of physics and become a physicist. Interest measures tell us nothing directly about abilities, though, as we shall see in a moment, there are certain relationships between abilities and interests. Interest measures and ability measures deal with two quite distinct aspects of fitness for a field of study or work. Each provides information that supplements

the other. Interest is not a substitute for ability, and, conversely, ability to learn the skills of a job is no guarantee of success or satisfaction in the job.

There have been many studies of the relationship between interest and ability.\* Most of these have related to aspects of academic work. In general, the relationship between achievement in a field such as science and the corresponding interest (i.e., scientific interest on the *Kuder*) is positive but low. Correlation of achievement with interest in the corresponding area will run about .30 to .50. Thus, there is some slight tendency for those with high ability for a field of knowledge to show high interest in it. But the relationship is much too low for either type of measure to serve in place of the other. Both types of information are needed for any sound evaluation of an individual's suitability for a particular program of study or plan for work.

Standardized interest inventories have been developed primarily for their contribution to vocational counseling and job placement. With this purpose in mind, they are directed at groups of high-school age or older. The *Kuder*, with its relatively general interest areas, has been used satisfactorily at about the ninth grade and above. The *Strong*, focusing on specific occupations and with a particular emphasis upon occupations at the professional level, is suitable primarily for senior high school pupils with definite plans to go to college and for college groups. As in almost all inventories, these instruments involve a good deal of reading. Their use with individuals who fall below eighth or ninth grade reading level would probably present serious problems.

Several other interest inventories are listed and briefly described in Appendix III.

## TEMPERAMENT AND ADJUSTMENT INVENTORIES

Self-report inventories have been extensively developed in the areas of temperament and personal adjustment. In these areas we again encounter instruments developed to yield scores for internally consistent clusters of behaviors, as did the *Kuder Preference Record*, and instruments built with keys based on reference to some external criterion, as was the *Strong Vocational Interest Blank*.

The basic material of all temperament and adjustment questionnaires is much the same. They draw from an extensive catalogue of statements about actions and feelings. To these the individual re-

\* See Frandsen \* for a review of some of this material.

sponds by indicating whether each is or is not characteristic of him. In many cases, a "?" or "uncertain" category is provided for the person who does not wish to endorse an unequivocal "Yes" or "No" answer. In the case of adjustment questionnaires, questions are culled from case studies, writings on various types of adjustment problems, suggestions of psychiatrists, and similar sources. For the normal dimensions of temperament, a review of psychological and literary treatments of personality differences and a systematic scrutiny of previous questionnaires, together with the personal insights of the investigator, provide the raw material for assembling items.

There are a large number of temperament and adjustment inventories. We will describe three in some detail, illustrating distinctively different patterns. These are the *Guilford-Zimmerman Temperament Survey*, the *Minnesota Multiphasic Personality Inventory (MMPI)*, and the *Edwards Personal Preference Schedule (EPPS)*. Then we will undertake a more general evaluation of the validity of inventories in this area and of the conditions under which we may expect them to be of value.

#### THE GUILFORD-ZIMMERMAN TEMPERAMENT SURVEY

The *Guilford-Zimmerman Temperament Survey* is the most recent development in a series of instruments on which Guilford has worked, each of which has attempted to identify and measure a number of internally coherent dimensions of personality that are clearly distinct from one another. Guilford has started with a pool of items and studied the intercorrelations among them, using the methods of factor analysis to which we referred on p. 262. He has identified distinct personality factors or foci, and tried to build up clusters of items to measure each. The objective is to get separate scales that are internally coherent and that are relatively independent of other scales. Thus, if a factor of "sociability" is identified, one attempts to get a cluster of items focussing on "sociability" that correlate substantially with each other, so that the person who subscribed to one item is likely also to subscribe to others. This cluster should be quite independent of other clusters relating to "dominance," "impulsiveness," and so forth, so that the correlations between the different clusters are quite low. This is the same basic approach as the one we saw in the *Kuder Preference Record*.

The *Guilford-Zimmerman* inventory provides scores appraising the clusters named and characterized below. Each cluster is characterized both by descriptive phrases and by two illustrative items.



*General Activity.* A high score indicates rapid pace of activities; energy, vitality; keeping in motion; production, efficiency, liking for speed; hurrying; quickness of action; enthusiasm, liveliness.

#### *Sample Items*

You start to work on a new project with a great deal of enthusiasm.

(+) You are the kind of person who is "on the go" all of the time. (+)

*Restraint.* A high score indicates serious-mindedness; deliberateness; persistent effort; self-control; *not* being happy-go-lucky or care-free; *not* seeking excitement.

#### *Sample Items*

You like to play practical jokes upon others. (-)

You sometimes find yourself "crossing bridges before you come to them." (+)

*Ascendancy.* A high score indicates habits of leadership; a tendency to take the initiative in speaking with others; liking for speaking in public; liking for persuading others; liking for being conspicuous; tendency to bluff; tendency to be self-defensive.

#### *Sample Items*

You can think of a good excuse when you need one. (+)

You avoid arguing over a price with a clerk or salesman. (-)

*Sociability.* A high score indicates one who has many friends and acquaintances; who seeks social contacts; who likes social activities; who likes the limelight; who enters into conversations; who is *not* shy.

#### *Sample Items*

You would dislike very much to work alone in some isolated place. (+)

Shyness keeps you from being as popular as you should be. (-)

*Emotional Stability.* A person with a high score shows evenness of moods, interests, etc.; optimism, cheerfulness; composure; feelings of being in good health; *freedom from* feelings of guilt, worry, or loneliness; *freedom from* day dreaming; *freedom from* perseveration of ideas and moods.

#### *Sample Items*

You sometimes feel "just miserable" for no good reason at all. (-)

You seldom give your past mistakes a second thought. (+)

*Objectivity.* The high scorer is defined as *free from* the following: egoism, self-centeredness; suspiciousness, fancying hostility; ideas of reference; a tendency to get into trouble; a tendency to be thin-skinned.

*Sample Items*

You nearly always receive all the credit that is coming to you for things you do. (+)

There are times when it seems everyone is against you. (-)

*Friendliness.* High scores signify respect for others; acceptance of domination; toleration of hostile action, *freedom from* hostility, resentment, or desire to dominate

*Sample Items*

When you resent the actions of anyone, you promptly tell him so. (-)

You would like to tell certain people a thing or two. (-)

*Thoughtfulness.* The high-scoring person is characterized as reflective, meditative; observing of his own behavior and that of others; interested in thinking; philosophically inclined; mentally poised.

*Sample Items*

You are frequently "lost in thought." (+)

You find it very interesting to watch people to see what they will do. (+)

*Personal Relations.* High scores signify tolerance of people; faith in social institutions; *freedom from* self-pity or suspicion of others.

*Sample Items*

There are far too many useless laws that hamper an individual's personal freedom. (-)

Nearly all people try to do the right thing when given a chance. (+)

*Masculinity.* The high-scoring person is interested in masculine activities; not easily disgusted; hardboiled; inhibited in emotional expression; resistant to fear; unconcerned about vermin; little interested in clothes, style, or romance.

*Sample Items*

You can look at snakes without shuddering. (+)

The sight of ragged or soiled fingernails is repulsive to you. (-)

Since each of these clusters can be thought of as a dimension having two ends, just as we have north and south, east and west, there is

an opposite end of each dimension that can be characterized as just the reverse of the description given above. Items marked (—) characterize this opposite end. Of course, most people do not score at either extreme on these dimensions. Here, as elsewhere, a continuous range of variation with most people occupying an intermediate position is the characteristic pattern. Most people are neither outstandingly active nor conspicuously lethargic, neither clearly ascendant nor clearly submissive. People can rarely be well described by clear-cut personality types. They are described as showing different *traits* in varying degrees.

Choosing the names for the clusters presented above was a bit of a problem, because the clusters do not correspond exactly to the language labels we bring with us. Each cluster is defined by the items that went into it and that were grouped together because they actually went together in the responses of people taking the inventory. The titles are approximate. Each cluster can be understood more exactly only by a close study of the items of which it is composed.

Table 12.5 \* shows the reliabilities of the separate scores, and the intercorrelations of the scores. The reliabilities cluster about .80 and are adequate, though not strikingly high. The attempt, in developing this inventory, was to identify a number of relatively independent aspects of personality. This means that the correlations of the different scores should be low. They tend to be. However, certain of the scores show rather substantial correlations. Attention may be directed to Ascendancy and Sociability, Emotional Stability and Objectivity, Friendliness and Personal Relations, and Restraint and Thoughtfulness. These pairs of scores are far from independent, and the information provided by the scores is overlapping. In a sense, the inventory is only partially efficient because of the duplication in the different scores. It is as if we were in part saying the same thing over

\* People who read about tests and testing will frequently have occasion to study tables of correlations like Table 12.5. In the table the column at the left lists the different variables and numbers them in order. The numbers (but not the names) are repeated across the top of the table. Look at the row labeled "1 General activity." The numbers that appear in this row are the correlations of "general activity" with each of the other variables. The first figure,  $-.16$ , is the correlation between "general activity" and variable 2, "restraint." This means that there is a slight tendency for high scores on the general activity scale to go with low scores on the restraint scale. The next figure,  $.34$ , is the correlation of "general activity" with "ascendancy," and the other entries are to be read in the same way. The correlation between any two variables will be found in the row and column whose numbers correspond to those variables. In this table, the reliability coefficients for the variables are shown in a column at the extreme right.

Table 12.5. Intercorrelations and Reliabilities of the Ten Scales of the *Guilford-Zimmerman Temperament Survey*

		Intercorrelations									Reliability *
Scale		2	3	4	5	6	7	8	9	10	
1	General activity	-.16	.34	.35	.34	.14	-.17	.24	-.03	.30	.79
2	Restraint		-.08	-.21	.08	.05	.25	.42	.14	-.01	.80
3	Ascendancy			.61	.35	.43	-.25	-.19	-.04	.29	.82
4	Sociability				.21	.36	-.06	.04	.18	.21	.87
5	Emotional stability					.69	.37	-.13	.34	.37	.84
6	Objectivity						.14	-.04	.43	.32	.75
7	Friendliness							-.05	.50	.26	.75
8	Thoughtfulness								.22	-.1*	.80
9	Personal relations									.35	.80
10	Masculinity										.85

\* Kuder-Richardson formula, based on 912 college students.

again. In most cases, however, each score provides information about a new and distinctive aspect of the individual.

The *Guilford-Zimmerman Inventory* has several characteristics that it may be well to summarize at this time.

1. It is based upon the responses of normal everyday people, not of the overtly maladjusted or the institutionalized.

2. Its scales are set up by internal analysis, by study of the "going together" of groups of items.

3. Responses are taken at face value. Their significance is assumed to be given by their obvious content.

4. The respondent may endorse as many or as few of the items as he wishes; his choices are not forced or constrained.

By contrast, let us consider the *Minnesota Multiphasic Personality Inventory*, which differs radically with respect to the three first features.

#### THE MINNESOTA MULTIPHASIC PERSONALITY INVENTORY

The *Minnesota Multiphasic Personality Inventory* was developed to identify a number of distinct categories of abnormal behavior. A pool of items was gathered which referred to different types of psychopathology: hysteria, depression, hypochondriasis, paranoid tendencies. The pool of items was tried out on a group of "normals" \* and upon

\* The problem of selecting a group of normal and well-adjusted persons is often a harder one than selecting people with a particular type of pathology. A particular type of disease can be identified with a good deal of definiteness, but absence of disease is a fuzzier notion, harder to define and to identify. The "normals" in this case were mostly people who had come to visit relatives at the Univ. of Minnesota Hospital.

a number of different groups with specific patterns of symptoms of maladjustment. The procedure was essentially the same as that for the *Strong*. Items were scored when they distinguished a given pathological group from the group of "normal" control cases.

The different scales of the *MMPI* are described below and illustrated with sample items. It must be remembered that the scales were established by using groups of patients showing behavior that was judged to be definitely abnormal. We cannot automatically apply the same labels to the variation in these traits that appears among normal individuals. The interpretation of scores found for normal persons must be made with great caution.

*Hypochondriasis Scale (Hs)*. This scale assesses the amount of abnormal or excessive concern with bodily functions. A high score indicates undue worry about health, often accompanied by reports of obscure pains and disorders that are difficult to identify.

#### *Sample Items*

I do not tire quickly. (-)

The top of my head sometimes feels tender. (+)

*Depression Scale (D)*. This scale appraises a tendency to be chronically depressed, to feel useless and unable to face the future.

#### *Sample Items*

I am easily awakened by noise. (+)

Everything is turning out just like the prophets of the Bible said it would. (+)

*Hysteria Scale (Hy)*. This scale gets at the tendency to solve personal problems by developing physical symptoms, such symptoms as paralyzes, cramps, gastric or intestinal complaints, or cardiac symptoms. The symptoms tend to appear under emotional stress and to be used as an escape mechanism.

#### *Sample Items*

I am likely not to speak to people until they speak to me. (+)

I get mad easily and then get over it soon. (+)

*Psychopathic Deviate Scale (Pd)*. This scale was based upon a group who showed absence of deep emotional response, inability to profit from experience, and disregard for social pressures and the regard of others. They were individuals who, from their disregard of social pressures, had tended to get into trouble of various sorts.

#### *Sample Items*

My family does not like the work I have chosen. (+)

What others think of me does not bother me. (+)

*Paranoia Scale (Pa).* The qualities evaluated by this scale are suspiciousness, oversensitivity, and feelings of being picked on or persecuted.

*Sample Items*

I am sure I am being talked about. (+)

Someone has control over my mind. (+)

*Psychasthenic Scale (Pt).* This scale was based on patients who were troubled with excessive fears or with compulsive tendencies to dwell on certain ideas or perform certain acts. High score indicates resemblance to this group.

*Sample Items*

I easily become impatient with people. (+)

I wish I could be as happy as others seem to be. (+)

*Schizophrenic Scale (Sc).* This scale is based upon a group of patients characterized by bizarre and unusual thought or behavior, and a subjective life tending to be divorced from the world of reality. High scores indicate responses similar to this group.

*Sample Items*

I have never been in love with anyone. (+)

I loved my mother. (-)

*Hypomania Scale (Ma).* This scale evaluates a tendency to be overactive both bodily and mentally, with a tendency to skip around rapidly from one thing to another.

*Sample Items*

I don't blame anyone for trying to grab everything he can get in this world. (+)

When I get bored I like to stir up some excitement. (+)

*Masculinity-Femininity Scale (Mf).* This scale measures interests characteristic of the one or the other sex.

*Sample Items*

I like movie love scenes. (F)

I used to keep a diary. (F)

The MMPI has a number of additional features, and these focus attention on certain problems that arise in using adjustment questionnaires. The first of these features is a *lie scale (L)*. This is based upon fifteen items, imbedded in the questionnaire, that relate to socially approved and virtuous activities that are generally approved of but not frequently carried out. General population norms indicate what

may reasonably be expected on a set of items of this sort. If a person marks an excessive number of these socially approved behaviors, it is considered to be an indication that he tends, consciously or unconsciously, to distort his report so that he appears in a favorable light. That is, he tends to "fake good."

Another score, the *K* scale, was built up by keying items that distinguished known abnormals who had presented normal score profiles from a control group of normals. A high score on this scale is thought to indicate a tendency to be very defensive in self-evaluation, whereas a low score brings out the tendency to be extremely self-critical, i.e., to "fake bad."

The ? score is based upon the number of ? or undecided responses. A very high number is thought to indicate a tendency to evade the task imposed by the inventory: to withdraw from it and fail to face up to it.

One further control scale is the *F* scale, made up of an assortment of unrelated items, each of which is marked as true only rarely in the general population. A high score on this scale is thought to be symptomatic of careless and superficial marking of the inventory: of marking items at random or misunderstanding the statements.

Thus, the authors of the *MMPI* have introduced a whole series of control scales, designed to isolate individuals whose responses are untrustworthy for one of several different reasons. They recognize, first, that good adjustment (and also bad adjustment) can be faked with at least partial success and that before an attempt is made to interpret scores on an inventory some guarantee is needed that there was not intentional faking. They recognize also that quite unintentionally individuals differ in the severity of the standards by which they judge themselves and that some control is needed on this difference in severity of standards. They recognize unwillingness to cooperate and inability to comprehend the task or to read the written items, which may show up as superficial and meaningless patterns of responses. All of these issues represent real problems to users of an inventory, and the control scores represent one well-conceived attempt to identify untrustworthy answer sheets.

In contrast with the *Guilford-Zimmerman*, we note that the *MMPI*:

1. Is based upon the distinctive responses of selected groups of persons—in this case, groups each presenting a particular psychopathology.
2. Has scales that are defined by these abnormal groups.
3. Is not concerned with the apparent meaning of an item, but only with whether it functions—whether it serves to differentiate between the abnormal and the control group.

It thus follows the general pattern of the *Strong Vocational Interest Blank*. In common with the *Guilford-Zimmerman*,

4. It permits any number of items to be endorsed, leaving the respondent free of constraint in this regard.

Let us look now at an inventory that makes use of the forced-choice pattern of response.

#### THE EDWARDS PERSONAL PREFERENCE SCHEDULE

The *Edwards Personal Preference Schedule* tries to assess the strength of various needs or motives in the life economy of the individual. Fifteen needs were selected from among those listed by Murray,<sup>11</sup> and items were developed to exemplify each. These are presented to the individual in pairs, each need being paired twice with each of the 14 others (to make a total of 210 items). Sample pairs are:

A I like to help my friends when they are in trouble.

B I like to do my very best in whatever I undertake.

A I like to conform to custom and to avoid doing things that people I respect might consider unconventional.

B I like to talk about my achievements.

The examinee must respond to each pair by indicating which statement is more true or more characteristic of him. Knowing how many times (out of 28) the examinee chose the option referring to achievement, for example, the examiner can refer to the norms and express *need-achievement* as a percentile of the norm group. Edwards made a systematic attempt to equate the statements in a given pair for *social desirability*, so that individuals would respond as they really felt, and not in terms of what is the approved or accepted thing to say. This was one way of trying to free scores of the element of defensiveness or "faking good" that has been a problem in many of the inventories that have been developed over the years.

The distinctive features of the *EPPS* are, then,

1. The "forced choice" pattern, which means that each respondent must make the same number of choices and the same number of rejections. Thus, no profile can be high on all scales, and each profile must have about the same number of highs and of lows. Everyone is brought to the same general base line. This is true also of the triads of the *Kuder*.

2. Equating "social desirability," so that any pressure or incentive to distort responses or "fake good" is held to a minimum.



## PROBLEM CHECK LISTS

The instruments we have just been describing yield one or several scores, representing traits or aspects of the individual. There are also several recently published "problem check lists," which are essentially catalogues of problems that are fairly often mentioned by children or young people. Examples are the *Mooney Problem Check List*, *SRA Junior Inventory* and *Youth Inventory*, and *Billett-Starr Youth Problems Inventory*. Responding to the comprehensive list of problems provides a kind of uniform problem-finding interview. The items that a child marks as matters of concern to him can serve as the starting point for more intensive inquiry in a face-to-face interview, while the problems that are marked as troublesome by several in a class group can serve as the focal point for group guidance sessions.

## EVALUATION OF TEMPERAMENT AND ADJUSTMENT INVENTORIES

How well can we hope to describe temperamental characteristics and personal adjustment through the individual's responses to a series of questions? Perhaps we can clarify the issue by asking what a person must do to fill out an inventory adequately. Completing one of these inventories usually requires that the respondent be (a) able to read and understand the item, (b) able to stand back and view his own behavior and decide whether the statement is or is not true of him, and (c) willing to give frank and honest answers. Each of these points raises certain issues about the validity of self-report instruments.

One problem in inventories of all types is that of reading load. This problem is partly one of sheer amount of reading. Especially in those inventories that try to appraise several different traits, it is usually necessary to have several hundred items to provide enough scope and reliability. The slow reader may have trouble getting through so much verbiage, or may start responding without really reading through the item. The problem is partly one of level of reading, i.e., of the complexity of structure and abstractness of ideas involved. If the vocabulary or concepts are beyond the respondent's comprehension, he may again give up the attempt really to understand and may respond in a superficial or random fashion. (The F scale of the *MMPI* was designed to protect against this hazard.) Thus, inventories are of questionable value for those of low literacy, be they adults or children.

A second problem is that of self-insight. Inventories require the individual to conceptualize and classify his own behavior—to decide whether certain descriptions or classifications of behavior are true of him. This implies a certain ability to stand back from himself and

view himself objectively that may be difficult to achieve. In fact, the person whose adjustment is most unsatisfactory may be the one who is least able to achieve this objectivity and to face his own deficiencies. Studies have shown repeatedly that those who are rated low by their associates on some desirable trait tend to grossly over-rate themselves. Thus, the ill-tempered girl is likely not to recognize her own irascibility; the overbearing boy may be unaware of his boorishness.

When inventories are built according to the pattern of the *Strong VIB* or the *MMPI*, such a lack of self-insight may not be of crucial importance. For these inventories, the keying of an item is based not on its obvious content but on the empirical fact that it did distinguish between criterion groups. If Henry has marked that he would like to be an architect, he has behaved in the way engineers typically behave. The question of whether engineers on the one hand or Henry on the other really want to be architects is not central to our interpretation. The point is that they have both reacted to the question in the same way, so we give Henry a credit on the engineer key of the *Strong*. On the other hand, where items and scores are interpreted on the basis of their manifest content and taken at face value, as is true of the *Guilford-Zimmerman* inventory or the *Kuder Preference Record*, non-insightful responses will lead to an untrue picture of the person who makes them.

A third problem is the willingness of the respondent to reveal the way he perceives or feels about himself. For personality inventories, frank and honest response by the examinee is essential for a valid picture. In most cases, the general significance of the items is reasonably apparent to the reader. Most subjects can follow successfully instructions to fake in a particular way. Even when the subject cannot fake successfully, if he tries to do so he will certainly give a distorted picture of himself. Inventory scores will only be useful when most respondents are answering in the way that they consider to represent themselves. The importance of providing protection against distortion is sufficiently great so that control scores to detect it have been introduced into the *MMPI* and certain other inventories.

This means that personality or adjustment inventories cannot be used, or can be used only with caution, when the examinee feels threatened by the test or feels that it may be used against him. Inventories have not generally proved useful in an employment situation, perhaps for this reason. If an inventory is given to elementary school pupils (and perhaps in high school and college) in the typical school setting, in which a test is something to do your best on and the teacher is often someone to get the best of, one is inclined to doubt whether many of

the pupils will be willing to reveal personal shortcomings that they may be aware of. Generally speaking, in any practical situation we should consider an adjustment inventory to be no more than a preliminary screening device that will locate a group of individuals who *may* be having problems of adjustment or *may* be in conflict with their environment. Final evaluation should always await a more personal and intensive study of the individual. Furthermore, a good score on an adjustment inventory is not a guarantee of good adjustment; it may characterize a person who is protective, defensive, or unable to face and to acknowledge very real problems.

Personality inventories are a product of the middle-class American culture. The extent to which items have equivalent meaning for other national cultures, or even for the lower socio-economic level in America, has not been fully explored. Some additional caution is necessary in interpreting results for members of other cultural groups.

*Evidences of Validity.* Those inventories that have been developed as measures of adjustment usually show a moderate level of *concurrent* validity. That is, they differentiate between groups established on other grounds as differing in adjustment. Thus, the *MMPI* was set up to distinguish between diagnosed pathological groups and normals and continues to do so in new groups. Other inventories have been tested by their ability to differentiate less extreme groups and have stood up fairly well under the test.

When it comes to *predictive* validity, the results are less encouraging. In civilian studies,<sup>1,2,7,17</sup> inventory scores have generally failed to predict anything much about the future success of the individual either in school, on the job, or in his personal living. Military experience<sup>3</sup> with these instruments has been somewhat more promising. There have been a number of studies showing substantial relationship between scores based on inventories and the subsequent judgment resulting from a psychiatric interview. Relationships to subsequent discharge from the service have also been sufficiently good to indicate that an inventory could serve a useful function as a device to screen for careful interview those who appeared to be potential misfits.

*The Practical Use of Temperament and Adjustment Inventories.* We must now ask what use should be made of temperament and adjustment inventories in and out of school. In the light of the factors that can distort scores and the limited validity these instruments have shown as predictors, we must conclude that they should be used very sparingly. Our feeling is that an adjustment inventory should be used only as an adjunct to more intensive psychological services. If facili-

tics are available to permit intensive study of some of the group by psychologically trained personnel, an inventory may serve as a means of identifying persons likely to profit from working with a counselor. However, there is little that a classroom teacher can do to dig behind and test the meaning of an inventory score. Accepted uncritically, the score may prove very misleading. We believe that little useful purpose is served by giving an adjustment inventory and making the results available to the teacher, especially the teacher of an elementary-school child.

The multi-dimensional temperament inventories are still too new for us to have much evidence on the social or practical validities of the different scales. Their use for vocational guidance or personnel selection can hardly be recommended at the present time. It may be that persons having certain patterns of temperamental characteristics should be guided towards or away from certain types of jobs. This seems plausible to many people. However, our information about the personality patterns in specific occupations is too limited, and the range of variation within occupations is probably too wide to make much practical use of such personality appraisals at the present time.

## ATTITUDE QUESTIONNAIRES

One further type of self-report inventory deserves brief mention. This is the attitude questionnaire, designed to appraise an individual's favorableness toward some group, proposed action, social institution, or social concept. Opinion polling has become commonplace in the last 25 years. However, this involves attitude *measurement* in only the most rudimentary sense. One or more questions are asked, and a count is made of the frequency of responses in two or three broad categories. Polls of this sort may be used in the schools to get an appraisal of public opinion of the school's patrons, or to study the status or change of pupils' expressed beliefs after instruction. The schools express a good deal of concern about development of attitudes and ideals as educational objectives, so there is need for good devices to appraise the extent to which such outcomes are being achieved. Industrial morale surveys are another point at which practical use may be made of attitude measurement. But the greatest use of attitude appraisal devices up to the present time has probably been for research studies of factors related to attitude differences, types of experiences that produce changes in attitude, or the influence of attitudes upon our perception of our world.

The typical attitude questionnaire is made up of a series of statements which the individual may either endorse or reject. There are two main patterns:

1. *Scaled Statements.* In this form, statements are scaled in terms of their degree of favorableness on the basis of extensive preliminary work. Thus, if we are preparing this type of attitude scale toward the United Nations, we start with a large pool of items. They may include the following:

The UN is a strong influence for peace.

The UN is a waste of time and effort.

The UN does about as much harm as good.

The UN is the most important force for good in the world today.

A corps of judges is assembled and each judge is asked to sort these statements into a set of piles, each pile representing a different degree of favorableness toward the UN. The judge is *not* indicating his agreement or disagreement with the statement; he is giving his interpretation of its meaning and significance. Each statement receives a scale value based on the average of these judgments and an ambiguity index based upon the spread of the ratings. (The more the judgments spread out, the more ambiguous the statement is.) From the pool of items tried out, about twenty are chosen that spread out over the range of scale values and are relatively unambiguous. These constitute the attitude scale.

When this type of attitude scale is administered, the respondent marks all the statements with which he agrees. His score is the average of the scale values of the statements he endorses.

2. *Summed Score.* In the other common format, the basic statements are much the same, except that neutral statements are avoided. Each statement is unequivocally either favorable or unfavorable. The respondent reacts to each statement on a five-point scale, ranging from strong agreement to strong disagreement. Thus, a section of a questionnaire in this format might read:

The UN is a strong influence for peace.	Strongly agree	Agree	Uncertain	Disagree	Strongly disagree
The UN will only make trouble.	Strongly agree	Agree	Uncertain	Disagree	Strongly disagree

The questionnaire can be scored quite simply by giving five points for strong endorsement of a favorable statement, four points for agreement, three points for uncertainty, and so forth. The scoring is reversed for the unfavorable statements. An individual's raw score is

the sum of his scores for the separate items. The raw score can, of course, be converted into a percentile or standard score if this seems desirable.

Both forms of attitude scale usually have satisfactory reliabilities, typically in the .80's. The two types of scales yield scores that intercorrelate very highly, and for most practical purposes there does not seem to be a great deal of choice between them. The greater simplicity of preparation of a summed-score type of inventory will commend it to most persons who wish to use an attitude scale as an aspect of some type of educational evaluation or research project. In either case, the scale will yield only a single general favorableness-unfavorableness score for an attitude area. Any qualitative variations within the broad area are blurred. Recent investigations on attitude scale development have been concerned with identifying more restricted and more homogeneous subscales within a larger attitude domain. A series of homogeneous subscales within a larger attitude area (toward the UN, for example) should permit mapping out in a more analytic and diagnostic way the profile of an individual's or a group's attitudes.

The big qualification about attitude scales is that they operate purely on a verbal level. The individual doesn't *do* anything to back up his stated attitude. The scales deal with verbalized attitudes rather than actions. Of course, an attitude scale is obviously fakeable. If we recognize that they represent the verbalized attitude that the individual is willing to express to us and work within that limitation, attitude scales appear to be a useful research tool or tool for experimental evaluation of educational objectives lying outside the domain of knowledge and skills.

## SUMMARY AND EVALUATION

In this chapter we have considered self-report inventories as instruments for studying personality. An inventory of this sort is essentially a standard set of interview questions presented in written form.

The individual's report about himself has one outstanding advantage. It provides an "inside" view, based on all the individual's experience with and knowledge about himself. However, self-reports are limited by the individual's limited

1. Ability to read the questions with understanding.
2. Self-insight and self-understanding.
3. Willingness to reveal himself frankly.

One type of questionnaire that has proven valuable in selection and placement is the biographical data blank, in which the individual provides factual information about his past history. A scoring key developed for the particular job has been found to have useful validity in several different instances.

Interest inventories provide satisfactorily reliable descriptions of interest patterns. These patterns persist with a good deal of stability, at least after late adolescence, and appear to be significant factors for vocational planning.

The validity of adjustment and temperament inventories is more open to question. Inventories of all types can be distorted to some extent if the individual is motivated to distort his responses. Thus, the integrity of the responses depends upon the motivation of the person examined. This depends, in turn, upon the setting in which and purposes for which the inventory is used. In school, industrial, or military use of adjustment inventories, one suspects that the motivations may often favor distorted responses. In any event, inventories of this type have not generally shown high validity. They should be used only with a good deal of circumspection.

Attitude questionnaires have been developed to score the intensity of favorable or unfavorable reaction to some group, institution, or issue. Though these represent only verbal expressions of attitude, they are useful research tools.

## REFERENCES

1. Ellis, A., Recent research with personality inventories, *J. consult. Psychol.*, 17, 1953, 45-49.
2. Ellis, A., The validity of personality questionnaires, *Psychol. Bull.*, 43, 1946, 385-440.
3. Ellis, A., and H. S. Conrad, The validity of personality inventories in military practice, *Psychol. Bull.*, 45, 1948, 385-426.
4. Fear, Richard, *The evaluation interview: prediction of job performance in business and industry*, New York, McGraw-Hill, 1958.
5. Frandsen, A., Interests and general educational development, *J. appl. Psychol.*, 31, 1947, 57-65.
6. Garry, R., Individual differences in ability to fake vocational interests, *J. appl. Psychol.*, 37, 1953, 33-37.
7. Ghiselli, E. E., and R. P. Barthol, The validity of personality inventories in the selection of employees, *J. appl. Psychol.*, 37, 1953, 18-20.
8. Longstaff, H. P., Fakability of the Strong Interest Blank and the Kuder Preference Record, *J. appl. Psychol.*, 32, 1948, 360-369.
9. Mallinson, G. G., and W. M. Crumrine, An investigation of the stability of interests of high school students, *J. educ. Res.*, 45, 1952, 369-383.

10. McCully, C. Harold, *The validity of the Kuder Preference Record*. Ed. D. Dissertation, George Washington University, Washington, D. C., 1954.
11. Murray, H. A., et al., *Explorations in personality*, New York, Oxford University Press, 1938.
12. Rosenberg, N., Stability and maturation of Kuder interest patterns during high school, *Educ psychol. Meas.*, 13, 1953, 449-458.
13. Strong, E. K., Interest scores while in college of occupations engaged in 20 years later, *Educ psychol. Meas.*, 11, 1951, 335-348.
14. Strong, E. K., Nineteen-year followup of engineer interests, *J. appl. Psychol.* 36, 1952, 65-74.
15. Strong, E. K., Permanence of interest scores over 22 years, *J. appl. Psychol.*, 35, 1951, 89-91.
16. Strong, E. K., *Vocational interests of men and women*, Stanford, Calif., Stanford University Press, 1943.
17. Super, D. E., The Bernscuter Personality Inventory: a review of research, *Psychol. Bull.*, 39, 1942, 94-125.

### SUGGESTED ADDITIONAL READING

- Allen, Robert M., *Personality assessment procedures*, New York, Harper, 1958, Chapters 2-7.
- Bass, Bernard M. and Irwin A. Berg, Editors, *Objective approaches to personality assessment*, New York, Van Nostrand, Chapters 1, 3, 5 and 6.
- Cronbach, Lee J., *Essentials of psychological testing*, 2nd ed., New York, Harper, 1960, Chapters 14-16.
- Darley, John G., and Theda Hagenah, *Vocational interest measurement*, Minneapolis, The University of Minnesota Press, 1955, Chapters 2, 4, 6.
- Harris, Chester W., Editor, *Encyclopedia of educational research*, 3rd ed., New York, Macmillan, 1960, pp. 102-112, 728-732.
- Guilford, J. P., *Personality*, New York, McGraw-Hill, 1959, Chapters 8-9.
- Kuder, Frederic G., *Kuder preference record occupational, Form D, research handbook*, 2nd ed., Chicago, Science Research Associates, 1957.
- Layton, Wilbur L., Editor, *The strong vocational interest blank: research and uses*, Minneapolis, University of Minnesota Press, 1960.

### QUESTIONS FOR DISCUSSION

1. How satisfactory is the method that was used in validating the *Strong Vocational Interest Blank*? What limitations do the procedures have? In what ways should they be checked?
2. What are the relative advantages of the *Strong Vocational Interest Blank* and the *Kuder Preference Record*? Under what circumstances would you choose to use one and under what circumstances the other?
3. What is the relationship between measures of interest and measures of ability? What does this suggest as to the ways in which the two types of tests should be used?



4. Most civilian studies have failed to find interest or adjustment inventories very useful in personnel selection. What are the reasons for this?

5. Why are most published interest inventories intended for use with secondary-school pupils, college students, and adults rather than elementary-school students?

6. What uses could a classroom teacher make of results on the *Kuder Interest Inventory* other than in giving vocational and educational guidance?

7. In what ways could a biographical data blank help a teacher in understanding the pupils in a class? What types of information would be useful to include on such a blank?

8. What conditions must be met if a self-report inventory is to be filled out accurately and give meaningful results?

9. How much trust can we place in adjustment inventories given in school to elementary-school children? What factors limit their value?

10. What important differences do you notice between the *Guilford-Zimmerman Temperament Survey* and the *Minnesota Multiphasic Personality Inventory*? For what purposes would each be more suitable?

11. What purposes are served by the control scales (*L*, *K*, *F*, ?) on the *Minnesota Multiphasic Personality Inventory*? What would be the comparable issues in personality rating scales? How might one adapt the ideas of control scales to ratings?

12. What factors limit the usefulness of paper-and-pencil attitude scales? What other methods might a teacher use to evaluate attitudes?

13. Prepare the rough draft for a brief attitude scale to measure teachers' attitudes towards objective tests.

14. With what kinds of groups can adjustment inventories be used most satisfactorily?

## Chapter 13



# The Individual as Others See Him

In the last chapter we considered the information about personality that could be gotten from inventories in which the individual describes himself. A second main way in which an individual's personality shows itself is through the impression he makes upon others. The second person serves as a reagent reacting to the first personality. How well does A like B? Does A consider B a pleasing person to have around? An effective worker? A good job risk? Does A consider B to be conscientious? Trustworthy? Emotionally stable? Questions of this sort are continually being asked of every teacher, supervisor, former employer, minister, or even friend. We must now inquire how fruitful it is to raise such questions and what precautions must be observed if the questions are to receive useful answers.

We shall first give brief consideration to the unstructured letter of recommendation. Then we shall examine rating scales and rating procedures. Finally, we shall consider some special forms of rating: nominating techniques and forced-choice rating procedures.

### LETTERS OF RECOMMENDATION

The most fluid form for getting an impression of one person through the eyes of a second person is to invite the second person to talk or write to you about him. Such a communication could be obtained in any setting. However, the setting in which it most commonly does occur is when person A is a candidate for something: admission to a school, a scholarship or fellowship, a job, membership in a club, or a security clearance. He then furnishes the institution, placement agency, or employer the names of people who know him well or know him in a particular capacity, and that agency obtains statements about A from B and C, who know him.

How useful and how informative is the material that is included in free, unstructured communications describing another person? Actu-

ally, in spite of the vast numbers of recommendations written every year, very little of a solid and factual nature is known about their adequacy or the effectiveness with which they discharge their function. Opinion covers the full gamut from a belief that a free and unconstrained letter about an applicant is the best possible way to get an evaluation of him to the conviction that letters of recommendation are completely worthless, from a conviction that the letter of recommendation is the core of any selection program to a feeling that the best thing to do with recommendations is to burn them. But factual studies of the reliability and validity of the information that is gotten from a letter of recommendation or of the extent to which recommendations influence the action taken with respect to an applicant are fragmentary in the extreme.

The letter of recommendation is such an unstructured document that it is very hard to study by sound research techniques. However, several investigators have attempted to make analyses of the content of the letters and to scale them with respect to the enthusiasm of the endorsement they provided. A moderate degree of agreement has been found<sup>4</sup> between different letters written about the same person. Within a group of applicants for jobs in secondary-school teaching from one teacher-training institution the between-letters reliability would be represented by a correlation of about .40. There was some evidence in this same study that the letters of those who got the jobs were a little higher on the enthusiasm scale than letters of applicants who were *not* employed. However, another study failed to find any difference between the terms used to describe job getters and other applicants.

The extent to which a letter of recommendation provides a *valid* appraisal of an individual and the extent to which it is accurately diagnostic of outstanding points, strengths or weaknesses, is almost completely unknown. However, we cannot be very sanguine. Most of the limitations that we shall presently discuss in connection with more structured rating scales apply with at least equal force to uncontrolled letters. In addition, each respondent is free to go off in whatever direction his fancy dictates, so that there is no core of content common to the different letters about a single person or to the letters dealing with different persons. One letter may deal with A's social charm; a second, with B's integrity; and a third, with C's originality. On what common base are we to compare the three? Add to this the facts that (1) the applicant usually is more or less free to select the persons who will write about him and may be expected to pick those who will support him and that (2) recommenders differ profoundly in their propensity for using superlatives, and the prospect is not a very rosy one.

Further research studies of the validity of free descriptions of one person by his fellows are urgently needed. In the meantime, recommendations will continue to be written—and perhaps to be used. We must turn our attention to more structured evaluation procedures.

## RATING SCALES

Undoubtedly it was in part the extreme subjectivity of the unstructured statement, the lack of a common core of content or standard of reference from person to person, and the extraordinary difficulty of quantifying the materials that gave impetus to the development of rating scales. Rating procedures attempt to overcome just these deficiencies. They attempt to get appraisals on a common set of attributes for all raters and ratees and to have these expressed on a common quantitative scale.

We all have had experience with ratings, either in making them or in having them made about us or, more probably, in both capacities. Rating scales appear in a large proportion of school report cards, more clearly in the non-academic part. Thus, we often find a section phrased somewhat as follows:

	1st Period	2nd Period	3rd Period	4th Period
Effort	_____	_____	_____	_____
Conduct	_____	_____	_____	_____
Citizenship	_____	_____	_____	_____
Cooperation	_____	_____	_____	_____
Adjustment	_____	_____	_____	_____

H = superior      S = satisfactory      U = unsatisfactory

Many civil service agencies and industrial firms send rating forms out to persons listed as references by job applicants, asking for evaluations of the individual's "initiative," "originality," "enthusiasm," or "ability to get along with people." These same companies or agencies often require supervisors to give merit ratings of their employees, rating them as "superior," "excellent," "very good," "good," "satisfactory" or "unsatisfactory" on a variety of traits or in over-all usefulness. Colleges, medical schools, fellowship programs, and still other agencies call for ratings as a part of their selection procedure. Beyond these practical operating uses, ratings have been involved in a great many research projects. All in all, vast numbers of ratings are called for and given, often reluctantly, in our country week by week and month by month. Rating other people is a large-scale operation.

The most common pattern of rating procedure presents the rater

with a set of trait names, perhaps somewhat further defined, and a range of numbers, adjectives, or descriptions that are to represent levels or degrees of possession of the traits. He is called upon to rate one or more persons on the trait or traits by assigning him or them the number, letter, adjective, or description that is judged to fit best. Two illustrations are given of rating scales, drawn from a program being developed for evaluation of management personnel.\* The first is one of a series of trait ratings. This part of the evaluation instrument calls for ratings of the following traits: job know-how, judgment, leadership, ability to plan and organize, communication ability, initiative, dependability, and human relations. For the trait of leadership, the rater is instructed as shown below. The actual rating scale follows these instructions.

### LEADERSHIP

Consider his ability to inspire confidence. How much respect does he command as an individual, not merely because of his position? Do people look to him for decisions? Is he afraid to "stick his neck out" for what he believes? Does he have teamwork?

Completely lacking. Definitely a follower with equals. Does not try to convince others that his way is best.

☐

Tries to lead with some success, but has never achieved a strong position. Is passive in directing his subordinates.

☐

Good leader. People wait to hear what he has to say. Respected by colleagues. People call for his opinion.

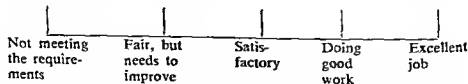
☐

Exceptional leader. Able to take over and pull things into shape. People seem to enjoy going along on his side. Is respected by subordinates and colleagues.

☐

An over-all summary rating is also called for, and this takes the form shown below.

Please place a mark on the scale to best show the over-all rating of this man in his present position.



\* These have been made available through the courtesy of the Personnel Department of Mack Trucks, Inc.

These are only illustrations of a wide range of rating instruments. We shall turn presently to some of the major variations in rating patterns. Right now, however, let us consider some of the problems that arise when we try to get a group of judges to make these appraisals.

## PROBLEMS IN OBTAINING SOUND RATINGS

The problems in obtaining valid appraisals of an individual through ratings are of two main sorts. There are first the factors that limit the rater's *willingness* to rate honestly and conscientiously, in accordance with the instructions given to him. There are secondly the factors that limit his *ability* to rate consistently and correctly, even with the best of intentions. We shall need to consider each of these in turn.

### FACTORS AFFECTING THE RATER'S WILLINGNESS TO RATE CONSCIENTIOUSLY

When ratings are collected, it is commonly assumed that each rater is trying his best to follow the instructions that have been given him, and that any shortcomings in his ratings are due entirely to human fallibility and ineptitude. However, this is not necessarily true. There are at least two sets of circumstances that may impair the integrity of a set of ratings: (1) The rater may be unwilling to take the trouble that is called for by the appraisal procedure; and (2) the rater may identify with the person rated to such an extent that he is unwilling to make a rating that will hurt him. Each of these merits some elaboration.

*Unwillingness to Take the Necessary Pains.* At best, ratings are a bother. Careful and thoughtful ratings are even more of a bother. In some rating procedures the attempt is made to get away from subjective impressions and superficial reaction by introducing elaborate procedures and precautions into the rating enterprise. Thus, in one attempt to improve efficiency rating procedures for Air Force officers,<sup>13</sup> an elaborate form was introduced that was to serve as a combined observational record and rating form. Fifty-four specific critical behaviors were described relating to officer efficiency. Scales were prepared describing degrees of excellence in each type of behavior. The accompanying instructions called upon raters to observe their ratees for a period before the official ratings were to be given and to tally on the rating form instances that had been observed of desirable and undesirable acts within each of the behavior categories described on the scale. After a year or two of use this form was discarded, in part at least because of its complexity and because raters were not willing to devote the time and thought that would have been required to maintain

the preliminary observational records on which the ratings were to be based.

In a lesser degree, one suspects that perfunctoriness in carrying out the operation of rating is a factor contributing to lowered effectiveness in many rating programs. Particularly if the number of pupils or employees to be rated is large, the task of preparing periodic ratings can become a decidedly onerous one. Unless raters are really "sold" on the importance of the ratings, the judgments are likely to be hurried and superficial ones, given more with an eye on finishing the task than with a concern for making accurate and analytical judgments.

*Identification with the Persons Being Rated.* Ratings are often called for by some rather remote and impersonal agency. The Civil Service Commission, the Military Personnel Division of a remote Headquarters, the personnel director of a large company, or the central administrative staff of a school system are all pretty far away from the first line supervisor, the squadron commander, or the classroom teacher. The rater is often closer to the persons being rated, the workers in his office, the junior officers in his outfit, the pupils in his class, than to the agency that requires the ratings to be made. One of the first principles of supervision or leadership is that the good leader looks out for the needs and welfare of his followers or subordinates. Morale in an organization depends upon the conviction that the leader of the organization will take care of the members of the group. When ratings come along, "taking care of" becomes a matter of seeing to it that one's own men fare as well as—or a little better than—those in competing groups.

All this boils down to the fact that in some situations the rater is more interested in providing a "break" for the people whom he is rating and in seeing that they get at least as good treatment as other groups than he is in providing accurate information for the using agency. This situation is aggravated in many governmental and official agencies by a policy of having the ratings public and requiring that the rater discuss with the person being rated any unfavorable material in the ratings. A further aggravation is produced by setting up administrative rulings in which a minimum rating is specified as required for promotion or pay increase. No wonder, then, that ratings tend to climb or to pile up at a single scale point. Thus, in certain governmental agencies during World War II the typical rating, accounting for a very large proportion of the ratings given, was "excellent." "Very good" became an expression of marked dissatisfaction, while a rating of "satisfactory" was reserved for someone you would get rid of at the first opportunity.

It is important to realize that a rater cannot always be depended upon to work wholeheartedly at giving valid ratings for the benefit of the using agency, that making ratings is usually a nuisance to him, and that he is often more committed to his own subordinates than to an outside agency. A rating program must be continuously "sold" and policed if it is to remain effective. And there are limits to the extent to which even an active campaign can overcome a rater's natural inertia and interest in his own little group.

#### FACTORS AFFECTING THE RATER'S ABILITY TO RATE ACCURATELY

Even when a group of raters are presumably well motivated and doing their best to provide valid judgments, there are still a number of factors that operate to limit the validity of those judgments. These center around the lack of opportunity to observe, the covertness of the attribute, ambiguity of the quality to be observed, lack of a uniform standard of reference, and specific rater biases and idiosyncrasies.

*Opportunity to Observe the Person Rated.* One factor that must always be borne in mind as a consideration limiting the accuracy of rating procedures is limited opportunity on the part of the rater to observe the person being rated. Thus, the high-school teacher teaching four or five different class groups of 30 pupils each and seeing many pupils only in a class setting may be called upon to make judgments as to the "initiative" or "flexibility" of these pupils. The college instructor who has taught a class of 100 pupils will receive rating forms from an employment agency or from the college administration asking for similar judgments. The truth of the matter is that effective contact with the person to be rated has probably been too limited to provide any adequate basis for the judgment that is being requested. True, the ratee has been physically in the presence of the rater for a good many hours, possibly several hundred, but these have been very busy hours, concerned primarily with other things than observing and forming judgments about pupil A. Pupil A has had to compete with pupils B, C, D, and on to Z and also with the primary concern with teaching rather than judging.

In a civil service or industrial setting much the same thing is true. The primary concern is with getting the job done, and although in theory the supervisor has had a good deal of time to observe each worker, in practice he has been busy with other things. We may be able to "sell" supervisors on the idea of devoting more of their energy to observing and evaluating the persons working for them, but there are very real limits to the amount of effort that can be withdrawn from a supervisor's other functions to be applied to this one.



We face not only the issue of general opportunity to observe, but also that of specific opportunity to observe a particular aspect of the individual's personality. Any person sees another only in certain limited contexts, in which only certain aspects of his behavior are displayed. The teacher sees a child primarily in the classroom, the foreman sees a workman primarily on the production line, and so forth.

We might question whether a teacher in a thoroughly conventional classroom has seen a child under circumstances which might be expected to bring out much "initiative" or "originality." The college instructor who has taught largely through lectures is hardly well situated to rate a student's "presence" or "ability to work with individuals." The supervisor of a clerk doing routine work is poorly situated to appraise "judgment." Whenever ratings are proposed, either for research purposes or as a basis for administrative actions, we should ask with respect to each trait being rated: Has the rater had a chance to observe these people in enough of the sorts of situations in which they could be expected to show variations in this trait so that his ratings can be expected to be meaningful? If the answer is "No," we would be well advised to abandon the ratings.

In this connection, it is worth while to point out that persons in different roles may see quite different aspects of the person to be rated. Her pupils see a teacher from quite a different vantage point than does the principal. Classmates in Officer Candidate School have a different view of the other potential officers than does the drill instructor. In getting ratings of some aspect of an individual, it is always appropriate to ask who has the best chance to see the relevant behavior displayed. It would normally be to this source that we should go for our ratings.

*Covertness of Trait Being Rated.* If a trait is to be appraised by an outsider, someone other than the person being rated, it must show on the outside. It must be something that has its impact on the outside world. Such characteristics as appearing at ease at social gatherings, having a pleasant speaking voice, and participating actively in group projects are characteristics that are essentially social. They appear in interaction with other persons and are directly observable. They are *overt* aspects of the person being appraised. By contrast, attributes such as "feeling of insecurity," "self-sufficiency," "tension," or "loneliness" are inner personal qualities. They are private aspects of personality and can only be crudely inferred from what the person does. They are *covert* aspects of the individual.

An attribute that is largely covert can be judged by the outsider only with great difficulty. Little of inner conflict or tension shows on the surface, and where it does show it is often in masquerade. Thus, a

child's deep insecurity may express itself as aggression against other pupils in one child, or as withdrawal into an inner world in another. The insecurity is not a simple dimension of overt behavior. It is an underlying dynamic factor that may break out in different ways in different persons or even in the same person at different times. Only a thorough knowledge of the individual, combined with a good deal of psychological insight, makes it possible to infer from the overt behavior the nature of his underlying covert dynamics.

One can see, then, that rating procedures will be relatively unsatisfactory for the inner, covert aspects of the individual. Qualities that depend upon very thorough understanding of a person plus wise inferences from his behavior will be rated with low reliability and little validity. Ratings have most chance of being accurate for those qualities that show outwardly as a person interacts with other people, the overt aspects. Experience has shown that these can be rated more reliably, and one feels confident that they are rated more validly. The validity lies in part in the fact that these social aspects of behavior have their meaning and definition primarily in the effects of one person or another.

*Ambiguity of Meaning of Dimension to Be Rated.* Many rating forms call for ratings of quite broad and abstract traits. Thus, in our illustration on p. 353 we included, among others, "citizenship" and "adjustment." These are neither more nor less vague and general than the attributes included in other rating schedules. But what do we mean by "citizenship" in an elementary-school pupil? By what actions is "good citizenship" shown? Does it mean not marking up the walls? Or not spitting on the floor? Or not pulling little girls' hair? Or bringing newspaper clippings to class? Or joining the Junior Red Cross? Or staying after school to help the teacher clean up the room? What does it mean? Probably no two raters would have just exactly the same things in mind when they rated a group of pupils on "citizenship."

Or consider "initiative," "personality," "supervisory ability," "mental flexibility," "executive influence," or "adaptability." These are all examples from rating scales in actual use. Though there is certainly some core of uniformity in the meaning that each of these terms will have for different raters, there is with equal certainty a good deal of variability in meaning from one rater to another. In proportion as a term becomes abstract, its meaning becomes variable from person to person, and such qualities as those listed above are conspicuously abstract.

The rating that a given child will receive for "citizenship" will, then, depend upon what "citizenship" means to the rater. If it means to

rater A conforming to school regulations, he will rate certain children high. If to rater B it means taking an active role in school projects, the high ratings may go to quite different children. A first problem in getting consistent ratings is to achieve consistency from rater to rater in the meanings of the qualities being rated.

*Uniform Standard of Reference.* A great many rating schedules call for judgments of the persons being rated in some set of categories such as

Outstanding, above average, average, below average, unsatisfactory.  
Superior, good, fair, poor.

Best, good, average, fair, poor.

Outstanding, superior, better than satisfactory, satisfactory, unsatisfactory.

Superior, excellent, very good, good, satisfactory, unsatisfactory.

But how good is "good"? Is a person who is "good" in "judgment" in the top tenth of the group with whom he is being compared? The top quarter? The top half? Or is he just *not* one of the bottom tenth? And what *is* the group with whom he is supposed to be compared? Is it all men of his age? All employees of the company? All men in his particular job? All men in his job with his length of experience? If the last, how is the rater supposed to know the level of judgment that is typical for men in a particular job with a particular level of experience?

The problem that all these questions are pointing up is that of forming a standard against which to appraise a given ratee. Variations in interpretation of terms and labels, variations in definition of the reference population, and variations in experience with the members of that background population all contribute to variability from rater to rater in their standards of rating. The phenomenon is a familiar one in academic grading practices. Practically every school that has studied the problem has found enormous variations among faculty members in the per cent of A's, B's, and C's that they give. The same situation holds for any set of categories, numbers, letters, or adjectives, that may be used. Standards of interpretation are highly subjective and vary widely from one rater to another. One man's "outstanding" is another man's "satisfactory."

Raters differ not only in the level of ratings that they assign, but also in how much they spread out their ratings. Some raters are conservative, and rarely rate anyone very high or very low; others tend to go to extremes. This difference in variability of ratings serves also to reduce the comparability of ratings from one rater to another.

*Specific Rater Idiosyncrasies.* Not only do raters differ in general "toughness" or "softness" They also differ in a host of specific idiosyncrasies. The experiences of life have built up in each of us an assortment of likes and dislikes and an assortment of individualized interpretations of the characteristics of people. You may distrust anyone who does not look at you while he is talking to you. Your neighbor may consider any man a sissy who has a voice pitched higher than usual. Your boss may consider that a firm handshake is the guarantee of a strong character. Your golf partner may be convinced that blonds are flighty. These are rather definite reactions that may be explicit and clearly verbalized by the person in question. But there are myriad other more vague and less tangible biases that we carry with us and that influence our ratings. These biases help to form our impression of a person and color all aspects of our reaction to him. They enter into our ratings too. In some cases, our rating of one or two traits may be affected. But often the bias is one of general liking for or aversion to the person, and this generalized reaction colors all our specific ratings. Thus, the ratings reflect not only the general subjective rating standard of the rater, but also his specific biases with respect to the person being rated.

#### THE OUTCOME OF FACTORS LIMITING RATING EFFECTIVENESS

What is the net result of these factors affecting the raters' willingness to rate conscientiously and ability to rate accurately? The effects show up in certain pervasive distortions of the ratings, in relatively low reliabilities, and in doubt as to the basic validity of rating procedures.

*The Generosity Error.* We have pointed out that the rater is often as much committed to the people he is rating as he is to the agency for which ratings are being prepared. Over and above this, there seems to be a widespread unwillingness on the part of raters to damn a fellow man with a low rating. The net result is that ratings tend quite generally to pile up at the high end of any scale. The unspoken philosophy of the rater seems to be "one man is as good as the next, if not a little better," so that "average" becomes in practice not the mid-point of a set of ratings but near the lower end of the group. One finds quite generally the paradox of a great majority of the group being rated above average.

If the generosity error operated uniformly for all raters, it would not be particularly disturbing. We would merely have to remember that ratings cannot be interpreted in terms of their verbal labels and that "average" means "low" and "very good" means "average." Makers of rating scales have countered this humane tendency with

some success by having several steps on their scale on the plus side of average, so that there is room for differentiation without having to get disagreeable and call a person "average."

It is differences between raters in the degree of their "generosity error" that are more troublesome. To correct for such differences is a good deal more of a problem. We shall consider presently some special techniques that have been developed for that purpose.

*The Halo Error.* Limitations in our experience with the person being rated, lack of opportunity to observe the specific qualities that are called for in the rating instrument, and the influence of personal biases that affect our general liking for the person all conspire to produce another type of error in our ratings. This is a tendency to rate in terms of over-all general impression without differentiating specific aspects, of allowing our total reaction to the person to color our judgment of each specific trait. This is called "halo."

We can illustrate halo by a set of data on embryo airplane commanders in World War II. Students were rated by their instructors for such qualities as "eagerness," "foresight," "leadership," "instrument flying," "formation flying," "lead crew potentiality," and "over-all value." The correlation between two raters for the same attribute was, on the average, about .60. This serves as a measure of the reliability of the ratings. We may speak of it as the between-raters reliability. The average correlation between *different* attributes for the *same* rater was about .75. That is, the correlation between ratings of different qualities was higher than the reliability of the separate ratings. This consistency can only be accounted for by a general halo that made instructor A's appraisal of student B much the same no matter what attribute was being rated.

Of course, some relationship among desirable traits is to be expected. We find correlation among different abilities when these are tested by objective tests and do not speak of the halo effect that produces a correlation between verbal and mechanical ability. Just how much of the relationship between the different qualities on which we get ratings is genuine and how much of it is spurious halo is very hard to determine. That some of the relationship is due to inability to free oneself from general biases seems clear, however, from examples such as the one we have just given.

*Reliability of Ratings.* Studies have shown repeatedly that the between-raters reliability of conventional rating procedures is low. Symonds,<sup>11</sup> writing in 1931, summarized a number of studies and concluded that the correlation between the ratings given by two independent raters for the conventional type of rating scale is about .55. There seems to be no good reason to change this conclusion after the lapse

of years. When the two ratings are uncontaminated; i.e., the raters have not talked over the persons to be rated, and where the usual type of numerical or graphic rating is used, the resulting appraisal shows only this very limited consistency from rater to rater.

If it is possible to pool the ratings of a number of independent raters who know the persons being rated about equally well, reliability of the appraisal can be substantially increased. Studies have shown<sup>25</sup> that pooling ratings functions in the same way as lengthening a test, and that the Spearman-Brown formula (p. 179) can legitimately be applied in estimating the reliability of pooled independent ratings. Thus, if the reliability of one rater is represented by a correlation of .55, we have the following estimates for the reliability of pooled ratings:

2 raters	.71
3 raters	.79
5 raters	.86
10 raters	.92

Unfortunately, in many important practical situations it is impossible to get additional equally qualified raters. An elementary-school pupil has only one regular classroom teacher; a worker has only one immediate supervisor. Adding on other raters who have limited acquaintance with the ratee may weaken rather than strengthen the ratings.

Reliability data on some of the newer types of rating devices to be discussed presently appear somewhat more promising. These data will be presented as the methods are discussed. One of the gains from basing ratings on specific tangible behaviors will be, it is hoped, that the objectivity, and hence the reliability, of the judgments will be increased.

*Validity of Ratings.* All the limiting and distorting factors that we have been considering make us doubtful about the validity of ratings. Rater biases and rater unreliability operate to lower validity. However, it is usually very difficult to make any statistical test of the validity of ratings. The very fact that we have fallen back on ratings usually means that no better measure of the quality in question is available to us. There is usually nothing else against which we can test the ratings.

In one context, the validity of ratings is axiomatic. If we are interested in appraising how a person is reacted to by other people, i.e., whether a child is well liked by his classmates or a foreman by his work crew, ratings are the reactions of these other persons and are directly relevant to the point at issue.

When ratings are being studied as predictors, statistical data can be obtained as to the accuracy with which they do in fact predict. This is something that must be determined in each setting and for each type of criterion that is being predicted. That ratings are in some cases the most valid available predictors is shown in studies of the ratings of aptitude for military service that are given at the U. S. Military Academy.<sup>12</sup> These ratings by tactical officers and by fellow cadets correlated more highly with later ratings of performance as an officer than did any other aspect of the man's record at West Point. Correlations with ratings of effectiveness in combat in the war in Korea were about .50. This criterion is again a rating, but it is probably as close to the real "pay off" as we are likely to get in this situation. In other situations, of course, ratings may turn out to have no validity at all. Each type of situation must be studied for its own sake.

### IMPROVING THE EFFECTIVENESS OF RATINGS

So far we have painted a rather gloomy picture of rating techniques as devices for appraising personality. It is certainly true that the hazards and pitfalls in rating procedures are many. But for all their limitations, there are and will continue to be a host of situations in which we will have to rely on the judgments of other people as a means of appraising our fellow men. The sincerity and integrity of a potential medical student, the social acceptability of a would-be salesman, the conscientiousness of a private secretary can probably only be evaluated through the judgment that someone makes of these qualities in the individuals in question. What can be done, then, to mitigate the defects of rating procedures? We shall consider first the design of the rating instrument and then the planning and conduct of the ratings.

#### REFINEMENTS IN THE RATING INSTRUMENT

The usual rating instrument has two main components: (1) a set of stimulus variables (the qualities to be rated) and (2) a pattern of response options (the ratings that can be given). In the simplest and most conventional rating forms, the stimulus variables consist of trait names and the response options consist of numerical or adjectival categories. Such a form was illustrated on p. 353. This type of format appears to encourage most of the shortcomings that we have been discussing in the preceding section. Consequently, many variations and refinements of format have been tried out in an attempt to overcome or at least minimize these shortcomings. The variations have manipu-

lated the stimulus variables, the response options, or both. Some of the main variations are described below.

#### REFINEMENTS IN PRESENTING THE STIMULUS VARIABLES

Bare trait names represent unsatisfactory stimuli for a rater for two reasons. In the first place, as we pointed out on p. 359, the words mean different things to different people. The child who shows "initiative" to teacher A may show "insubordination" to teacher B, whereas teacher B's "good citizen" may seem to teacher A a "docile conformist." In the second place, the terms are quite abstract and far removed in many cases from the realm of observable behavior. Consider "adjustment," for example. We do not observe a child's adjustment. We observe a host of reactions to situations and people. Some of these reactions are perhaps symptomatic of poor adjustment. But the judgment about the child's adjustment is several steps removed from what we have a chance to observe.

Workers with ratings have striven to get greater uniformity of meaning in the traits to be rated, and they have attempted to base the ratings more closely upon observable behavior. These attempts have modified the stimulus aspect of rating instruments in three ways.

1. *Trait Names Have Been Defined.* A phrase, sentence, or several sentences have been appended to each trait name to give it greater uniformity of meaning. Thus, we might have:

*Citizenship.* Participation in school projects. Willingness to do his share. Responsibility for work and property.

This represents a somewhat more objective and behavioral statement and should produce at least *some* more uniformity in meaning among a group of raters. However, we may doubt that a brief verbal definition will completely overcome the individual differences in meaning that different raters bring to the task.

2. *Trait Names Have Been Replaced by Several More Concrete and Limited Descriptive Phrases.* Thus, the abstract and blanket term "citizenship" might be broken down into the several components suggested above, i.e.:

Participation in school projects.  
Willingness to do his share.  
Responsibility for completing work.  
Carefulness with school property.

A judgment would now be called for with respect to each of the more limited and more concrete aspects of pupil behavior.



3. *Each Trait Name Has Been Replaced by a Substantial Number of Descriptions of Specific Behaviors.* This carries the move toward concreteness and specificity one step farther. Following out our analysis of "citizenship," we might replace it with a set of behaviors somewhat as follows:

- a. Works well with other children in groups and committees.
- b. Brings materials to school.
- c. Does his work without complaining.
- d. Gets assigned work in on time.
- e. Keeps desk and work area neat.
- f. Uses materials without wasting.
- g. Works steadily, even when not watched.
- h. When one task is done, finds other work to do.
- i. Takes care of school property.

This list is still more tangible and specific. There should be relatively little opportunity, in each case, for ambiguity as to what it is that is being observed and reported on.

The replacement of one general term with many specific behaviors gives promise of achieving more uniformity of meaning from one rater to another. It may also bring the ratings in closer touch with actual observations that have been made of the behavior of the individual who is being appraised. Where the trait to be rated is one that the rater has really had no opportunity to observe, the attempt to replace the trait name with specific observable behaviors will often make this fact painfully apparent and will force the designer of the instrument to rethink the problem of relating his instrument to the observations that the rater has really had an opportunity to make.

The gains that a list of specific behaviors achieves in uniformity of meaning and concreteness of behavior judged are not without cost. The cost lies in the greatly increased length and complexity of the rating instrument. There are limits to the number of different judgments that can be asked of a rater. Furthermore, the lengthy, analytical report of behavior may be confusing to the person who tries to use and interpret it. The lengthy list of specific behaviors will probably prove most effective when (1) judgments are in very simple terms, such as simply present-absent and (2) there are provisions for organizing and summarizing the specific judgments into one or more scores for broad areas.

#### REFINEMENTS IN FORM OF RESPONSE CATEGORIES

Expressing judgments about a ratee by selecting some one of a set of numbers, letters, or adjectives is still common on school report cards



of this sort may also be used to present the choice alternatives. Thus, we may have an item of this type:

*Participation in School Projects*

Volunteers to bring in materials. Suggests ideas. Often works overtime.	Works or brings materials as requested. Participates, but takes no initiative.	Does as little as possible. Resists attempts to get him to help.
---	--	--

In this case, three statements describing behavior are combined with a graphic scale, and are used to define three points on the scale. The descriptions may be expected to lend more concreteness and uniformity of meaning to the scale steps. However, these editorial provisions do not completely overcome rater idiosyncrasies, which continue to plague us.

4. *Man-to-Man Scales.* An early attempt to get more uniformity of meaning into the response scale, developed in World War I, used men instead of numbers, adjectives, or descriptions to represent the scale points. The rater is asked to think of someone he has known well who was very high on the quality being rated. That person's name is then entered on the rating form to define the "very high" point on the scale. In the same way, the names of other persons known well by the rater are entered in spaces to define "high," "average," "low," and "very low." The five names then define levels for the trait. When a person is to be rated, the rater is instructed to compare him with the five persons defining the levels on the trait. The rater is to judge which man he most closely resembles on the trait in question. He is assigned the value corresponding to the step on the scale which that man occupies.

It was thought that the man-to-man feature would lend concreteness to the comparisons and overcome the tendency of some raters to be consistently generous. In cases in which all raters have a wide range of acquaintance, so that their scale persons may be expected to be fairly comparable, the procedure may make for more uniformity from rater to rater. But such scope of acquaintance and thoroughness of familiarity with suitable scale persons is likely to be somewhat unusual in the practical situations in which ratings must be made. Implicit comparison with other persons is involved in any rating enterprise, but explicit use of particular persons to define the steps on a rating scale has not been widely adopted.

5. *Present—Absent.* When a large number of specific behavioral statements are used as the stimuli, the response that is called for is

often a mere checking of those that apply to the individual in question. The person is then characterized by the statements that are checked as representing him. The rating scale becomes a behavior check list. The set of items on p. 366 might constitute part of such a check list.

If this type of appraisal procedure is to yield a score, the statements must be scaled or assigned score values in some way. The simplest way is merely to score them +1, -1, or 0, depending upon whether they are favorable, unfavorable, or neutral with respect to a particular attribute (i.e., perseverance, integrity, reliability, etc.) or a particular criterion (i.e., success in academic work, success on a job, responsiveness to therapy, etc.). An individual's score can then be the sum of the scores for the items checked for him.

If the additional elegance seems justified, more refined scaling procedures can be applied to the statements. Scale values can be based on their judged significance or the degree to which they had actually discriminated between successful and unsuccessful individuals. The score an individual receives is then based on an averaging of the scale values of the items that were checked as describing him. The reliability of such a check list of scaled items has been found to be quite satisfactory in some instances, Richardson and Kuder<sup>14</sup> reporting a correlation of .83 between two independent raters of groups of salesmen.

Only limited use has been made of check lists as devices to yield scores on each individual, but they seem to present a promising pattern. They come the closest of any of the rating procedures to self-report inventories on the one hand and to ability tests on the other. A behavior check list is in a sense a personality inventory that has been filled out by someone other than the person being described. The items can be selected and scored in much the same way. The resemblance to an ability test can be seen in one well-known behavior check list, the *Vineland Social Maturity Scale*.<sup>3</sup> This check list is made up of items relating to self-help, self-direction, communication, socialization, and the like. Selected items from different levels of the scale are shown in Table 13.1.

Norms for the scale were established for each item, representing the age at which the behavior appears on the average. The check list is filled out by a rater who knows the child being appraised. Items the person does or can do are checked. A basal age is established for which all items are positive, and the person being rated is automatically given credit for all earlier items. Points are given for additional items passed. The table of norms gives developmental age equivalents for

Table 13.1. Items Selected from the Vineland Social Maturity Scale

Item No.	Age Level (in years)	Item
1	0-1	"Crows," laughs
6	0-1	Reaches for nearby objects
11	0-1	Drinks from cup assisted
15	0-1	Stands alone
19	1-2	Marks with pencil or crayon
28	1-2	Eats with spoon
34	1-2	Talks in short sentences
37	2-3	Removes coat or dress
40	2-3	Dries own hands
44	2-3	Relates experiences
51	4-5	Cares for self at toilet
53	4-5	Goes about neighborhood unattended
68	7-8	Disavows literal Santa Claus
70	7-8	Combs or brushes hair
78	10-11	Writes occasional short letters
80	10-11	Does small remunerative work

the point scores, and a developmental quotient may be computed that indicates the individual's rate of progress toward self-sufficiency and independence.

The check-list pattern has been used as a simple descriptive instrument, as in school reports to the home. The procedure is attractive in this setting because it can give information on specific aspects of pupil development. However, forms tend to become complicated and to confuse many parents, so this type of reporting has not been widely adopted.

6. *Frequency of Occurrence, or Typicality.* Instead of reacting in an all-or-none fashion to an item, as in the check list, response can be qualified as being "always," "usually," "sometimes," "seldom," or "never" characteristic of the ratee. Or the ratee may be characterized as "very much like," "a good deal like," "somewhat like," "slightly like," or "not at all like" the behavior described in the statement. (The terms of frequency or resemblance may vary; the ones given are only suggestive.) An individual's score would now take account both

of the significance of the statement and the point on the scale that was checked. That is, an important attribute would receive heavier credit than a minor one, and a check at the "always" step more credit than a check at "usually."

Indefinite designations of frequency or degree of the sort that are being discussed here will be differently interpreted by different raters, so the old problem of differences in rater standards is still with us. Moreover, when the number of specific behaviors being checked is substantial, a simple present-absent checking correlates quite highly with the more elaborate form.

7. *Ranking.* In those cases in which each rater knows a substantial number of ratees, he may be asked to place them in rank order with respect to each attribute being studied. Thus, a teacher may be asked to indicate the child who is most outstanding for contributing to the class projects and activities "over and beyond the call of duty," the one who is second, and so on. Usually, the ranker will be instructed to start at both ends and work in toward the middle, since the extreme cases are usually easier to discriminate than the large group of average ones in the middle. In order to ease the task of the ranker, tie ranks may be permitted. If no tie ranks are permitted, the ranker may feel that the task is an unreasonable one, especially in a group of some size.

Ranking is an arduous task for the ranker, but it does achieve two important objectives. It forces the person doing the evaluation to make discriminations among those being evaluated. The ranker cannot place all or most of the persons being judged in a single category, as may happen with other reporting systems. Secondly, it washes out individual differences among raters in generosity or leniency. No matter how kindly the ranker may feel, he must put somebody last, and no matter how hardboiled he is, someone must come first. Individual differences in standards of judgment are eliminated from the final score.

If scores based on rankings by different judges are to be combined, there is one assumption that is introduced in rankings that may be about as troublesome as the individual differences in judging standards that have been eliminated. If we are to treat rankings by different judges as comparable scores, we must assume that the quality of the group ranked by each was the same. That is, we assume that being second in a group of twenty represents the same level on the trait being appraised, whichever group of twenty it happened to be. Usually we do not have any direct way of comparing the different subgroups, so about all we can do is assume that they are comparable. If the groups are fairly sizable and chosen more or less at random from the

same sort of population, this may be a reasonable assumption. But with small groups or groups selected in different ways, the assumption of comparability may introduce substantial amounts of error into any scores based on ranks.

Ranks as such do not represent a very useful score scale. The meaning depends upon the size of the group: being third in a group of three is very different from being third in a group of thirty. Furthermore, steps of rank do not represent equal units of a trait. As we saw in our discussion of percentile norms (Chapter 6), in the usual bell-shaped distribution, one or two ranks at the extremes of a group represent much more of a difference than the same number of ranks near the middle of the group. For that reason, it is common practice to convert ranks into normalized standard scores in order to get a type of score that has uniform meaning without regard to the size of the group and uniform units throughout the score range. Special tables have been prepared to facilitate this conversion, and tables for groups of all sizes up to twenty-five may be found on pp. 90-92 of Symonds (ref. 17).

#### THE "FORCEO-CHOICE" PATTERN

All the variations considered so far operated on the same basic pattern. The rater considered one attribute at a time and assigned the ratee to one of a set of categories or placed him relative to others on that particular attribute. We shall now consider a major departure from that pattern. The essence of the procedure we consider now is that the rater considers a *set* of attributes at one time and decides which one (or ones) most accurately represents the person being rated. Thus, an instrument developed for evaluating Air Force technical-school instructors<sup>3</sup> included sets of items such as the following:

- a. Patient with slow learners.
- b. Lectures with confidence.
- c. Keeps interest and attention of class.
- d. Acquaints classes with objective for each lesson.

The rater's assignment was to pick out the two items from the set that were *most descriptive* of the person being rated.

Note that all the statements in the above set are nice things to say about an instructor. As a matter of fact, they were carefully matched, on the basis of information from a preliminary investigation, to be just about equally nice to say about an instructor. But they differ a good deal, again based on preliminary investigations, in the extent to which they actually distinguish between persons who have been identified on other evidence as being good and poor instructors. The

most discriminating statement is (a) and the least discriminating is (b). Thus, we could assign a score value of 2 to statement (a), 1 to (c) and (d), and 0 to (b). A person's score for the set would be the sum of the credits for the two items marked as most descriptive of him. His score for the whole instrument would be the sum of his scores for 25 or 30 such blocks of four statements. Such a score was found to have good split-half reliability (.85 to .90), so that this instrument provided a reliable score for the individual's desirability as an instructor in the eyes of a single rater. This does not, of course, tell anything about the agreement that would be found between different raters.

By casting the evaluation instrument into a forced-choice format, the maker hopes to accomplish three things:

1. He hopes to eliminate variation in rater standards of generosity or kindness. Since the items in a set are all equally favorable things to say about a person, the kindly soul should have no particular tendency to choose one rather than another, and the true nature of the ratee should be the controlling factor.

2. He hopes to minimize the possibility of a rater intentionally biasing the score. In the ordinary rating scale, the rater is in pretty complete control of the situation. He can rate a man up or down as he pleases. In the forced-choice type of instrument, it is hoped that the rater will be unable to identify which are the significant choices and that therefore he will be unable to throw the score one way or the other at will. However, though there are some indications that a forced-choice instrument is less fakeable than an ordinary rating scale, it is still far from tamper-proof in the hands of a determined rater.

3. He hopes to produce a better spread of scores and a more nearly normal distribution of ratings. By making all options equally attractive, one minimizes the effect of the generosity error, it is hoped, and gets a more symmetrical spread of scores. Again, there is indication that this result is achieved at least in part.

Forced-choice rating instruments are a relatively new development, dating from World War II, though the forced selection of one of a set of alternates had been used before that time in self-report inventories. The close similarity in the pattern of these forced-choice ratings to self-report instruments such as the *Kuder Preference Record* and the *Edwards Personal Preference Schedule* should be apparent. Because of the relative novelty of the forced-choice pattern, evaluation of its usefulness in merit rating procedures and in personality appraisal is still incomplete. This format does appear to get away from some of the most troublesome limitations of conventional rating procedures.



However, it has some limitations of its own.<sup>1</sup> It has a tendency to create rater resistance, because of the difficulty of the judgments that the rater is called upon to make. Where the options are negative, i.e., "Is this worker more stupid or more lazy?" the instrument has a good deal of the "Have you stopped beating your wife yet?" flavor. And even the judgment as to whether employee A is more intelligent or more industrious is not easy to make. There often seems to be no basis for comparing two quite different traits.<sup>2</sup> The score that results from this type of instrument does not have any clear trait label or psychological interpretation, even if it is a relatively good predictor of some particular criterion. It gives us little help in building a descriptive picture and an understanding of the individual.

Developmental and exploratory work with forced-choice rating instruments continues. For example, a recent version produced in the Standard Oil Company of New Jersey as a Management Performance Report combines forced-choice with numerical rating. A set of four items would appear as follows:

	Fits poorly									Fits well								
Follows work schedule closely	0	1	2	3	4	5	6	7	8	9								
Has good work habits	0	1	2	3	4	5	6	7	8	9								
Is a credit to his department	0	1	2	3	4	5	6	7	8	9								
Makes decisions promptly	0	1	2	3	4	5	6	7	8	9								

The numerical scale runs from a low of 0 to a high of 9. The rater may use any part of the scale, with the one restriction that he may not use the same scale point for two statements. Thus, he can rate a man relatively low on all or relatively high on all. This takes some of the onus out of the forced ranking so far as the rater is concerned. In using the results, we may treat them either as conventional ratings, paying attention to the level checked, or as pure forced-choice rankings, ignoring the numerical values completely.

#### REFINEMENTS IN THE RATING PROCEDURES

The best-designed instrument cannot give good results if used under unsatisfactory rating conditions. Raters cannot give information they do not have and cannot be made to give information they are unwilling to give. We must, therefore, try to pick raters who have had close contacts with the ratees and ask them for judgments on attributes they have had an opportunity to observe. We should give them some guidance and training in the type of judgments we expect them to make, and if possible they should have opportunity to observe the ratees *after*

they have been educated in the use of the ratings. When there are several people who know the *ratees* equally well, ratings should be gathered from all of them and pooled. Every effort should be made to motivate the raters to do an honest and conscientious job. Let us consider these points further.

*Selection of Raters.* For most purposes, the ideal rater is the person who has had a great deal of opportunity to observe the person being rated in situations in which he would be likely to show the qualities on which ratings are desired. (Occasionally it may be desirable to get a rating of the impression which a person makes on brief contact or in a limited experimental situation.) It is also desirable that the rater take an impartial attitude toward the *ratee*. The desirability of these two qualities, thorough acquaintance and impartiality, is generally recognized in the abstract. However, the goals may be only partially realized in practice.

Administrative considerations usually dictate that the rating and evaluation function be assigned to the teacher in the school setting and to the supervisor in a work setting. The relationship here is in each case one of direct supervision. There is generally a continuing and fairly close personal relationship. But the relationship is a one-directional and partial one. The teacher or supervisor sees only one side of the pupil or worker, the side that is turned toward the "boss."

Those qualities that a boss has a good chance to see, primarily qualities of work performance, can probably be rated adequately by the teacher or supervisor. Thus, in one study<sup>2</sup> of airplane mechanics it was found that the ratings by a pair of supervisors on "job know-how" were as reliable as the pooled ratings by eight coworkers in a plane maintenance crew and that the supervisors' pooled rating correlated .53 with a written proficiency test, whereas the pooled rating for the coworkers correlated only .43. However, those qualities that show themselves primarily in relationships with peers or subordinates will probably be evaluated more soundly by those same peers and subordinates. The validity of the U. S. Military Academy peer ratings described on p. 364 is a case in point.

The lack of agreement between supervisor and pupil ratings of teachers is suggested in some of the following correlations from different studies:

Pupil's rating of excellence versus principal's rating <sup>2</sup>	.39
Pupil's rating of excellence versus composite of 5 judges <sup>2</sup>	.28
Mean pupil rating of effectiveness versus administrator's rating <sup>1</sup>	.08
Student versus administrator rating on general teacher effectiveness <sup>14</sup>	.40
School I	
School II	.50

A certain amount of overlap does exist, but the ratings appear also to have a good deal of uniqueness. The bird's eye and worm's eye views are not the same.

*Who Should Choose the Raters?* The selection of persons to rate applicants for jobs or fellowships requires consideration from another point of view. In this setting, the applicant is usually asked to supply a certain number of references or to submit evaluation forms filled out by a certain number of individuals. The choice of the individuals is usually left up to him, and we may anticipate that he will select persons he believes will rate him favorably. It might be more satisfactory if the applicant were asked to supply the names and addresses of persons who stood in particular relationships to him and who should be able to supply relevant information, rather than leaving the applicant free to pick his own endorsers. Thus, a job applicant might be asked to give the names of his immediate supervisors in his most recent jobs; a fellowship applicant, to list the name of his major advisor and of any instructors with whom he had taken two or more courses. Thus, we are shifting the responsibility of determining who shall provide the ratings from the applicant to the using agency. Such a shift should reduce the amount of special pleading for the applicant.

*Selection of Qualities to Be Rated.* Two principles appear to apply in determining the types of information to be sought by rating procedures. In the first place, it seems undesirable to use rating procedures to get information that can be provided satisfactorily by some more objective and reliable indicator. Score on a well-constructed intelligence test is a better indicator of intellectual ability than a supervisor's rating of intellect. When accurate production records exist, they are to be preferred to a supervisor's rating for productivity. Ratings are something to which we resort when we do not have any better indicator available.

Secondly, we should limit ratings to relatively overt qualities, ones that can be expressed in terms of actual observable behavior. We cannot expect the rater to look inside the ratee and tell us what goes on within. Furthermore, we must bear in mind the extent and nature of the contact between rater and person rated. For example, a set of ratings to be used after a single interview should be limited to the qualities that can be observed in an interview. The interviewee's neatness, composure, manner of speech, and fluency in answering questions are qualities that are observable in a single interview. His industry, integrity, initiative, and ingenuity are not, though these qualities might be appraised with some accuracy by the person who has worked with him for a time. Ratings should be of observable behavior—observable in the setting in which the man has been observed.

*Educational Program for Raters.* Good ratings do not just happen, even with the proper raters and the proper instrument for recording the ratings. Raters must be "sold" on the importance of making good ratings and taught how to use the rating instrument. Pointing out the importance of "selling" a rating program is easier than telling how to do it. As we have indicated earlier, inertia on the one hand and identification with the ratee on the other are powerful competing motives. We cannot provide a course in direct selling at this point, but a job of selling needs to be done in almost any program for gathering ratings. Furthermore, the selling must continue if thoughtfulness and integrity of the appraisals are to be maintained.

It is desirable that raters have practice with the specific rating instrument. A training session, in which the instrument is used under supervision, is often desirable. The meanings of the attributes can be discussed, sample rating sheets can be prepared, and the resulting ratings reviewed. The prevailing generosity error can be noted, and raters cautioned to avoid it. Further practice can be given, in an attempt to generate a more symmetrical distribution of ratings. Training sessions will not eliminate all the shortcomings of ratings, but they should reduce somewhat the more common distortions considered earlier.

*Observations Made as a Basis for Ratings.* One objection to ratings is that they are usually made after the fact and are based on general unanalyzed impressions about the person rated. An attempt to get away from this dependence on general memory is sometimes made by introducing the rating program well in advance of the time at which the final ratings are to be called for. It is hoped that the raters will then be on the alert for and take specific note of behavior relating to the qualities that are to be rated. As noted on p. 354, the attempt has even been made to provide for systematic recording of such observations over a period of time. However, recording of this type calls for a high level of commitment to, and cooperation in, the rating program. Where that level of involvement is achieved, advance notice and systematic recording may be expected to improve the rating process. Situations of this sort are probably rare, however.

*Pooling of Ratings by Several Raters.* One of the limitations of ratings is low reliability. In those situations in which there are a number of persons who have all had approximately equal chance to observe the ratee, it may be possible to get independent ratings from each potential rater and to pool these into a composite rating. Studies have shown<sup>12</sup> that the effect on reliability of pooling independent ratings is essentially the same as the effect of lengthening a test. The formula given in Chapter 7 (p. 187) applies. Thus, theoretically we could

achieve any needed level of reliability in our appraisal merely by increasing the number of raters.

The catch is found in the phrase "equal chance to observe the ratee." Unfortunately, the number of persons well placed to observe a person in some particular setting, school, job, camp, etc., is usually limited. Often only one person has been in close contact with the ratee in a particular relationship. He has had only one homeroom teacher, only one foreman, only one tent counselor. Others have had some contact with him, but it may be so much less that their judgments add little to the judgment of the rater most intimately involved.

Note that we specified the pooling of *independent* ratings. If the ratings are independently made, the "error" components will be independent and will tend to cancel out. If, however, the judgments are combined through some sort of conference procedure, we cannot tell just what may happen. Errors may cancel out, wisdom may win, or the prejudices of the most dogmatic may prevail. Pooling independent judgments is the only sure way of balancing out individual errors and has been found in several studies<sup>10,11</sup> to be more satisfactory than the conference type of procedure.

#### NOMINATING TECHNIQUES

If a teacher is to understand pupils, he must have some awareness of the values and standards that the group sets for its members—the peer culture—and of the role that each child plays in the group of his contemporaries—the peer group. The standards and values of his peers provide the sanctions and the rewards that are very influential in determining how a person will act and how content he will be in the group setting. The peer group can be quite a cohesive unit. In such a group any action by a teacher with respect to an individual child is often viewed not only as an action for or against him but also as an action for or against the group to which he belongs and which identifies with him. Thus, in order both to understand the individual and to understand how acts with respect to individuals affect the group climate, it is important to appraise the role of the individual in the group.

It is far from easy for the teacher or other outsider to get an accurate appraisal of group structure and of the place of the individual in it. The child's role is likely to be seen only from an adult point of view and that adult viewpoint to be projected upon the group of his contemporaries. Thus, when a child is helpful, friendly, and generally acceptable to the teacher, the teacher is likely to attribute to that child a level of influence with other children that he does not have. It is

often difficult for the teacher to attribute to an active and troublesome child his true level of influence with his peers. Teachers are often only dimly aware of the pattern of social interplay in their classroom, the reputation of each pupil among his peers, the factors determining prestige in the peer group, the patterns of attraction and repulsion, or the individual social aspirations.

In the understanding of these relationships, peer ratings are often helpful. A rating procedure that is very simple and quite effective for obtaining appraisals by peers is the *nominating technique*. We will consider this technique first as applied to social choices and rejections and then as applied more generally to trait ratings.

To improve their understanding of the social structure in a classroom, the patterns of friendship and leadership, teachers may use the simple expedient of asking pupils to name their choices of best friends or of work partners. For example, a teacher might say to a class: "For our unit on Mexico, we are going to need some committees of children who will work together on some part of the project. I would like to know which children you would like to have on a committee with you. Put your name on the top of the piece of paper I gave you. Then under it put the names of the children you would especially like to have on your committee."

We now have a series of nominations or choices for work partners. It is possible to show these choices pictorially by a diagram such as that shown in Fig. 13.1. This is called a *sociogram* and the procedure of constructing a sociogram is called *sociometry*.

Procedures to help in the construction of sociograms can be found in Moreno<sup>3</sup> and in a booklet by the staff of the Horace Mann Lincoln Institute.<sup>4</sup>

From the sociogram shown in Fig. 13.1, we see that A and B are the most sought after members of the group: these are the "stars." Pupils J and O did not choose anyone and were not chosen by any other pupils: they are isolates. Pupils H and I chose each other but were not chosen by any other pupils. Except for the mutual friendship between them, they too are isolates. Pupils P, Q, M, and N are fringers: they do not really belong to any of the groups but do make choices within the group.

Figure 13.1 shows the pattern of choices and attractions within the group. It would also be possible to have children indicate those class members whom they would definitely *not* want in their group. Calling for rejections presents some slight risks to individual and class morale but does permit a more complete picture of group structure.

The sociogram in Fig. 13.1 indicates that this is not a closely-knit

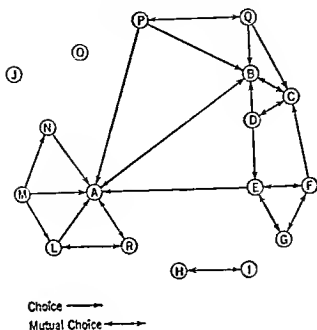


Fig. 13.1. Sociogram of fourth-grade class.

group. The rather large number of isolates and fringers and the linkages across from one "clique" to the other suggest an unstable pattern which is in the process of changing and reforming. Thus, the sociogram might represent a class at the beginning of the school year, in which a residue of last year's friendships is mixed with new currents and in which pupils from other class groups and other schools are not yet integrated into the group. It is in such a setting as this that the teacher can be most effective in bringing isolates into the group or promoting new friendships.

After the teacher has determined which children are without friends or are relatively isolated in the group, he should try to find out why this is the case. Sometimes the explanation may be very simple. The child may be new to the group and have not yet had time to find his place in it. The normal opportunities to get acquainted, furthered by the teacher's efforts to bring out the new child's assets, may be all that is required. The child may be older or younger than the rest of the group, having friends in other classes or outside of school. The child may not live near any of the other children in the class. At other times, the reasons may be more subtle, and it may take a good deal of discreet sleuthing for the teacher to find out why Willie or Alice are not chosen by their classmates.

When the reasons are understood, the teacher can often help to

remove them. Sometimes the simple process of coaching the child so that he develops competence in athletics may turn the trick. The teacher can arrange seats so that a child is placed near one for whom he expressed preference. Sometimes helping a child to develop everyday social graces or to improve his personal appearance is all that is needed to make him acceptable. If an isolate or fringer has special mechanical or artistic skills, giving him an opportunity to use these in class group activities may be effective.

In general, the teacher can help a child become integrated with and accepted by his peer group by (1) providing opportunity for developing friendly relations, (2) improving social skills, and (3) building up a sense of accomplishment or competence.

Sociometric choices describe the present flow of interaction among children rather than indicating any strong and permanent emotional structuring. However, the structuring of a class group affects the general emotional climate of the classroom. In a class where there are many isolates or children who are "fringers," i.e., not completely accepted by a clique, the morale of the group tends to be low and group planning and coordinated group action is made more difficult. It is also true that the teacher in dealing with one child is quite frequently dealing with the clique to which the child belongs.

Sociograms frequently point up mistakes that a teacher makes in characterizing a child. Thus, when the teacher has judged a child and his position in his peer group by adult standards, sociometric devices point out these mistakes and give the teacher a framework for understanding behavior that taken by itself may seem unexplainable.

Sociograms have been used in various non-school situations. In industry they have been used to form work groups and have been found to stimulate production. They have been used in institutions, especially those for juvenile offenders, to select house groups.

The sociogram by itself tells the teacher only what children are selected or rejected, not the reasons for selection and rejection. It is most useful when used in conjunction with good anecdotal records. For successful use, especially when rejections are asked for, there needs to be a friendly feeling between the teacher and the class. Furthermore, the teacher should actually use the nominations as far as possible in the way in which he has told the class he would use them.

The teacher should also remember that group structure is not static, especially in younger age groups. One sociogram made at the beginning of a school year will rarely provide an adequate picture of group structure through the year. Furthermore, neither choices nor rejections can be taken entirely at face value. When, as is sometimes the



procedure, the number of choices is limited to "three best friends," failure to choose a particular pupil need not mean lack of friendly feeling for him. Choices may reflect the prestige of the person chosen and a desire to be associated with that prestige, rather than a link of friendship. The culture pattern in certain age groups dictates that rejections follow sex lines. Class and caste distinctions also introduce cultural factors influencing choice and rejection. A sociogram is at best a rough and tentative picture of the social currents and climate of the group.

A final word of caution should be sounded about attempting to use sociometric data to reconstruct a group or modify a child's role in it. We have offered some suggestions as to ways in which a teacher may try to help the relatively isolated child. However, any such manipulations call for a good deal of subtlety. Heavy-handed attempts by the teacher to manipulate the pupils in the group may only aggravate the ills he is trying to cure.

Other patterns for obtaining peer evaluations have been developed, and they have been used for other purposes beside the preparing of sociograms and the studying of social currents within the group. A slightly more complex form is the *Ohio Social Acceptance Scale*, in which each pupil reacts to each other pupil in the group, checking him under one of the following six categories: (1) My very, very best friends, (2) My other friends, (3) Not friends, but okay, (4) Don't know them, (5) Don't care for them, (6) Dislike them. From the pooled pupil responses, a score may be obtained for each child indicating the extent of his acceptance within the group. This or some other similar format provides a simple procedure for obtaining ratings by a group of peers, and their simplicity makes them usable even with elementary school children.

Nominations may be used at any age level, and may be made with respect to any type of characteristic. For example, they have frequently been used in the armed services in Officer Candidate School, where each member of a unit may be asked to nominate a specified number of individuals in his unit who have shown the greatest evidence of "leadership" during the training course. He may also be asked to nominate those who have shown the *least* indication of leadership.

Taking all the nominations for the group as a whole, it is possible to arrive at a score for each individual, giving a plus for each favorable nomination and a minus for each unfavorable nomination.

A variation of the nominating procedure that has been used with school children has usually been referred to as the "Guess Who" tech-

nique or as "Casting Characters." In this procedure, the children are instructed somewhat as follows:

Suppose we were going to put on a class play. The characters in the play are described below. For each character, you are to put down the names of one or more children in the class who would be good for that part because he or she is just like that anyway.

"This person is always cheerful and happy—never grouchy or cross.

"This person is always butting in and telling other people how to do things. He cannot mind his own business

"This person is very quiet and doesn't get into games or do things with other children."

The number of characters can be extended as desired. Each "character" is a description in fairly concrete terms of a quality of behavior in which the investigator is interested. Descriptions of opposite ends of a scale can be included—i.e., friendly versus unfriendly, dominating versus submissive, etc.—and can be treated as positive and negative nominations on a single scale. Each child receives a score for each "character," based on the number of nominations he receives.

The attractive feature of the nominating pattern is its simplicity, which makes it rather painless to administer and usable with young groups or groups with little sophistication or experience in rating. It is feasible because the large number of raters make it possible to use a simple count of nominations instead of a rating of the usual type.

## SUMMARY AND EVALUATION

In spite of all their limitations, evaluations of persons through ratings will undoubtedly continue to be widely used for administrative evaluations in schools, civil service, and industry, as well as in educational and psychological research. We must recognize this fact and learn to live with it. Granting that we shall continue to use ratings of different aspects of personality, we should do so with full awareness of the limitations of our instruments, and we should do so in such a way that these limitations are minimized.

The limitations of rating procedures arise out of:

1. A humane unwillingness to make unfavorable judgments of our fellows, which is particularly pronounced when we identify to some extent with the person being rated (generosity error).

2. Wide individual differences among raters in "humaneness" or, in any event, in leniency or severity of rating (differences in rater standards).

3. A tendency to respond to other persons as a whole in terms of our general liking or aversion and difficulty in differentiating out specific aspects of the individual personality (halo error).

4. Limited contact between the rater and person being rated—limited both in amount and in type of situation in which seen.

5. Ambiguity in meaning of the attributes to be appraised.

6. The covert and unobservable nature of many of the inner aspects of personality dynamics.

7. Instability and unreliability of human judgment.

In view of these limitations it is suggested that ratings will provide a most accurate portrayal of the person being rated when:

1. Appraisal is limited to those qualities that appear overtly in interpersonal relations.

2. The qualities to be appraised are analyzed into concrete and relatively specific aspects of behavior, and judgments are made of these behaviors.

3. A rating form is developed that forces the rater to discriminate and/or that has controls for rater differences in judging standards.

4. Raters are used who have had the most opportunity to observe the individual in situations in which he would display the qualities to be rated.

5. Raters are "sold" on the value of the ratings and trained in the use of the rating instrument.

6. Independent ratings of several raters are pooled when there are several persons qualified to carry out ratings.

Evaluation procedures in which the significance of his ratings is somewhat concealed from the rater present an interesting possibility for civil service and industrial use. This is true particularly when controls on rater bias are introduced through "forced-choice" techniques or a correction score.

Peer-nominating techniques have interesting possibilities for use in schools and other group settings. They permit sociometric analyses of the interpersonal relations of pupils in a classroom or the workers in a shop. "Guess Who" nominations permit a simple type of rating in the early grades.

## REFERENCES

1. Brookover, W. B., Person-person interaction between teachers and pupils and teaching effectiveness. *J. educ. Res.*, 34, 1940, 272-287.
2. Cook, W., and C. H. Leeds, Measuring the teaching personality. *Educ. psychol. Meas.*, 7, 1947, 399-410.

Guilford, J. P., *Psychometric methods*, 2nd ed., New York, McGraw-Hill, 1954, Chapter 11.

Harris, Chester W., Editor, *Encyclopedia of educational research*, 3rd ed., New York, Macmillan, 1960, pp. 809-812, 929-931, 1320-1322.

## QUESTIONS FOR DISCUSSION

1. If you were writing to someone who had been given as a reference by an applicant for a job in your company or for admission to your school, what should you do in order to obtain the most useful evaluation of the applicant?
2. Make as complete a list as you can of the different ratings used in the school that you are attending or the school in which you teach. What type of a rating scale or form is used in each case?
3. In the light of such evidence or opinion as you can obtain, how effective are the ratings that you identified in the previous question? How adequate a spread of ratings is obtained? How consistently is the scale used by different users? What is your impression of the reliability of the ratings? Of their freedom from halo and other errors?
4. What factors influence a rater's willingness to rate conscientiously? How serious is this issue? What can be done about it?
5. Why would three *independent* ratings from separate raters ordinarily be preferable to a rating prepared by the three persons working together as a committee?
6. In the personnel office of a large company, employment interviewers are called upon to rate job applicants at the end of the interview. Which of the following characteristics would you expect to be rated reasonably reliably? Why?
  - a. Initiative.
  - b. Appearance.
  - c. Work background.
  - d. Dependability.
  - e. Emotional balance.
7. In a small survey of the report cards used in a number of communities the following four traits were most frequently mentioned as found on the report cards: (a) courteous, (b) cooperative, (c) health habits, (d) works with others. How might these be broken down or revised so that the classroom teacher could evaluate them better?
8. Which of the following would influence your judgment of a person in an interview? In what way?
  - a. A very firm grip in shaking hands.
  - b. Wearing a "loud" necktie.
  - c. Generally pausing for a moment before replying to a question.
  - d. Playing with keys on a key ring.
  - e. Having a spot on his vest.
  - f. Looking at the floor all during the interview.

9. Compare the reactions of several class members or of several acquaintances on the items of question 8. How general are the reactions? What basis in fact is there for them?

10. What advantages do ratings by peers have over ratings by superiors? What disadvantages?

11. What are the advantages of ranking over rating on a rating scale? What are the disadvantages?

12. Suppose that a forced-choice rating scale had been developed for use in rating the teachers in a city school system in order to get an evaluation of their effectiveness. What advantages would this rating procedure have over other types of ratings? What problems would be likely to arise in using it?

13. Make up a "Guess Who" form that might be useful to a teacher in finding out about the pupils in his class. If a class group is available to you, try the form out and analyze the results. What precautions should be taken in using the results?

14. Using a class group taught by some class member or made available by the instructor, get each child's choices for other children to work on a committee with him. Plot the results in a sociogram. What do the results tell you about the class and the pupils in it? What limitations would this sociogram have for judging the status of an individual child among his classmates?

15. Suppose you have been placed in charge of a merit rating plan which is being introduced in some company. What steps would you take to try to get as good ratings as possible?

## Chapter 14



# Behavioral Measures of Personality

We have tended to define personality as the typical quality of an individual's behavior. It would be natural, then, to go directly to the behavior of the individual to get an appraisal of his personality. Two possibilities are available to us. We may set up especially designed "test" situations, in which the individual's behavior may be scored or rated. Or we may plan to observe his behavior as it occurs spontaneously in his natural environment. Each of these has received attention from psychologists and educators, and we shall consider each in turn.

### BEHAVIOR TESTS

In personality testing we are concerned with the typical behavior of the individual—what he *will* do under the ordinary conditions of life, rather than what he *can* do if he is trying to do his best. Under these circumstances, it is obvious that any test must usually be indirect and disguised, so that the examinee does not know what is being appraised. This appears especially clearly in the field of character testing.

Traits of character relate to behaviors in which society sets up definitions of what is "good" and what is "bad." We can hardly expect a child to report his dishonesties, for example, or to show them in a test situation in which he knows his honesty is being observed and appraised. Furthermore, he has probably managed to conceal most of his transgressions from teacher, camp counselor, or other adult who might be asked to rate him. We are almost forced back upon a concealed test to elicit such socially disapproved behavior. We shall describe in some detail the honesty tests devised by May and Hartshorne for the Character Education Inquiry,<sup>8</sup> in part for their intrinsic interest and in part because they illustrate the virtues and many of the limitations of this type of measurement procedure.

May and Hartshorne developed a comprehensive series of tests of honesty. These included situations in which the individual had a chance to cheat, situations in which he had an opportunity to lie, and situations in which it was possible for him to steal. Some of the situations are described below.

*Situation A: Cheating on a test by copying.* A test is given dealing with some topic related to school work, word knowledge, for example. The papers are collected. The next day the papers are passed out, and each pupil is allowed to score his own paper when the answers are read aloud. As a matter of fact, however, the papers have been accurately scored before they are returned without any marks being made on the paper. The amount that the pupil copies in and scores his own paper above the correct score is used as an indication of cheating.

*Situation B: Cheating on a test by adding on.* A speeded arithmetic test is given, and at the end of 2 or 3 minutes pupils are told to stop work. However, for several minutes papers are left on their desks while the teacher or test administrator is busy doing something else. Later a second test is given after which the papers are immediately collected. When performance on the first testing surpasses performance on the second test by a specified amount, this is taken as evidence that the examinee added onto his work after the time limit was up and before the papers were collected.

*Situation C: Cheating in a game—peeking.* The game is illustrated in Fig. 14.1. The stunt is to shut one's eyes and put a dot in each

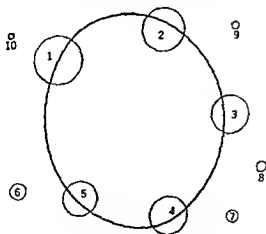


Fig. 14.1. Aiming test. (After Hartshorne and May.<sup>3</sup>)

circle in turn. Norms are prepared, based upon children tested with their view blocked so that they cannot peek. A child who performs unduly well, as determined by the "peek-proof" norms, is assumed to have peeked and helped himself.

*Situation D: Cheating in an athletic contest.* As a part of a "field day," each child is given a hand dynamometer to squeeze as a test of strength of hand. Three "practice" squeezes are given, and the adult observer notes and later records the best performance on these. Then the pupil is told to make additional squeezes "for the record." While he makes the squeezes, the adult is obviously busy with another child and not watching him. The child records his own performance on a record blank. Since fatigue tends to set in on successive squeezes, it is unlikely that he will show improvement. If the performance he reports surpasses his practice squeezes by a specified amount, it is assumed that he has been unduly optimistic in recording his performance.

*Situation E: Lying—self-glorification.* In this test the child is asked a series of questions. Each question has to do with standards of behavior that are universally applauded but seldom achieved. Thus, one question reads "Do you always obey your parents cheerfully and promptly?" and another, "Do you always smile when things go wrong?" It is hard to know how many of a set of statements like this a child might truthfully endorse, but an attempt was made to determine this by having groups of graduate students think back to their childhood and respond as would have been true of them then. The child who marks an excessive number of items is deemed to be not angelic but untruthful.

*Situation F: Stealing.* A game is devised which uses a number of coins. These are in a box, and one box is passed out to each child. After the game is over, each child is told to put the coins back in the box and fasten it up. The boxes are collected. They have been unobtrusively coded, so it is possible to tell which child had which box. A check of the coins in the boxes makes it possible to determine which children have helped themselves to one or more of the coins.

As can be seen from the brief descriptions, the tests are quite involved and require rather extensive stage-managing. The details of the testing situation seem fairly critical, i.e., how sure the child feels that he is free from observation, the manner in which the children are occupied when they are stopped in their work, and so forth. And it is crucial that the "security" of the test be maintained, for if the true purpose of the test were suspected, examinees could immediately conform to the approved social standard.



## EVALUATION OF BEHAVIOR TESTS OF HONESTY

How reliable and how valid are these situational tests of honesty? Reliability estimates are shown in Table 14.1. We can see that the reliabilities of single tests are rather modest, averaging about .50. In comparison with the aptitude and achievement tests we have been considering in the preceding chapters, these reliabilities are disappointing. The score of a pupil on any single test of the set used by May and Hartshorne would provide only the roughest indication of the typical behavior of that child. A single test would need to be extended by adding on several additional tests of the same sort if a satisfactorily stable and dependable measure were to be obtained. The tests would appear to be useful primarily for the comparison of different groups of pupils.

Table 14.1. Reliabilities of Tests Used for Measuring Deception  
(From May and Hartshorne<sup>6</sup>)

Type of Test	Reliability Coefficient
1. Copying from a key or answer sheet	.70
2. Adding onto one's score on a speeded test	.44
3. Peeping when one's eyes should be shut	.46
4. Faking a solution to a puzzle	.50
5. Faking a score in a physical ability test	.46
6. Lying to win approval	.84
7. Getting illicit help at home	.24

When it comes to validity, we are put to it to find any outside standard against which to evaluate the tests. Teachers' ratings of pupils may be taken as one limited and imperfect criterion, and the classroom cheating tests showed a modest correlation with this criterion (average about .35). But before we look for outside criterion measures, we should perhaps ask how the different kinds of honesty tests correlate with each other.

Considering four different types of cheating tests carried out in the classroom situation, the authors found that on the average a test of one type correlated with a test of one of the other types only to the extent of .26. When some type of classroom cheating test was correlated with cheating in an athletic contest, the average correlation was found to be only .16, and with the stealing test the average correlation was .17. The lying test, also given in the classroom, averaged .23 with the other classroom tests and .06 with the two out-of-classroom tests.

Even though the reliabilities of the single tests are low, the correlations between the different sorts of tests are a good deal lower. When the correlations involve different settings (i.e., classroom versus gymnasium) or different types of behavior (i.e., cheating versus stealing), the correlations drop still further. Many of them are not far from zero. The common factor running through all these tests is of quite limited significance, and to a very considerable extent the different tests are measuring different specific factors.

These results can be thought of as telling us something about the "trait" of honesty (and probably about the nature of other personality and character traits), and something about performance testing of personality. So far as the general trait is concerned, it is of limited significance in determining and predicting behavior in a specific situation. Behavior is determined to a very considerable extent by the specifics of the situation, and by specific factors (habits or response patterns) in the individual that have developed in response to that situation. Our characterization of a person in general terms will be only partially effective in predicting what he will do in a given specific setting.

So far as personality testing is concerned, we find that each performance test behaves much like an *item* on an ability test or a personality inventory. Thus, in a sense, the test that permits a pupil to revise his answer sheet as he scores it really asks a specific question, in behavioral terms, to wit: Would you change the answers if you scored your own paper? Each honesty test asks a somewhat different question, and collectively they might be thought of as a twenty-item questionnaire on honesty. The reliabilities of the separate tests, and their intercorrelations are not so different from those that we find for the single items of a personality questionnaire. If we think of the single tests as items, we will appreciate better the relatively modest reliabilities and the quite considerable specificity that they show.

The low correlations among specific honesty tests make it necessary to include a number of separate tests if we hope to get an adequate representation of different honesties. Because of this fact, together with the complexities of testing procedure, the use of behavior tests of character has been limited largely to research projects. They have not been adapted to any extent for routine use in schools or for any type of personnel selection.

#### SELECTED RESULTS FROM MAY-HARTSHORNE STUDIES

As research tools, behavior tests have provided a wealth of interesting data, notably in the original studies of May and Hartshorne. Some

of the more interesting findings from these studies are summarized below. Readers are referred to the original studies for details.

1. Honesty was essentially unrelated to age or sex over the range of grades studied. There was no tendency for children to learn to be more honest as they got older.

2. The more intelligent children received higher honesty scores. Of course, school pressures were probably less severe for brighter children. How much the difference in behavior reflects a difference in motivational pressures cannot be determined.

3. Honesty was associated with socio-economic status, children from higher socio-economic levels evidencing less dishonesty than those from lower levels.

4. Siblings resembled one another in honesty, and this resemblance was more than could be accounted for by familial resemblances in intelligence or by the common socio-economic background.

5. Children in a school following progressive educational practices cheated less than comparable children in a conventional school program.

6. The children within a school as a whole or a class group within a school tended to resemble one another in level of honesty displayed. There appeared to be a factor of school or class morale.

7. There was no indication that children who participated in organized programs of religious education or who were members of groups expressing character education aims were more honest than non-participants or non-members.

#### OTHER TYPES OF PERFORMANCE TESTS

A number of psychologists have recently been exploring indirect performance measures as indicators of personality variables. Eysenck<sup>4</sup> has developed a battery of performance measures to predict neuroticism. Some of the measures that have tended to discriminate between normal and neurotic groups are (1) the amount of body sway in response to a direct suggestion of falling, (2) the number of unusual responses on a multiple-choice free association test, (3) the speed of dark adaptation, (4) the number of food aversions, and (5) the length of time breath could be held. From a battery of eight or ten such specific tests Eysenck was able to get quite high reliability and fairly sharp discrimination between a normal and a neurotic group. Cattell<sup>5</sup> has attempted to develop batteries of performance measures to appraise the same personality factors that he had identified in personality rating and inventory procedures. These approaches are ap-

pealing, in that they presumably cannot be distorted by the subject to give a desired impression. However, the procedures are complex and time-consuming, and the value of the tests is still largely undetermined. Indirect, objective testing to appraise personality remains primarily an area for further research.

## SITUATIONAL TESTS AND ASSESSMENT PROGRAMS

During and since World War II a number of assessment programs have been set up for making a comprehensive appraisal of candidates for a particular type of training or assignment. Perhaps the most publicized of these was the program set up to screen personnel for the Office of Strategic Services during World War II. The program has been fully described,<sup>1</sup> and some features of it will be worth considering here. Assessment programs have generally made use of a wide variety of techniques for evaluating the individual. They have included ability tests of several sorts, detailed interviews, and various types of fantasy and projective materials. However, one central element has been the situational test, in which the individual is placed in a more or less standardized task situation where his behavior can be observed, his responses recorded, or various aspects of his reactions rated by observers.

### SITUATIONAL TESTS IN THE OSS ASSESSMENT PROGRAM

For assessment by the OSS staff, each candidate was brought to an assessment center for a 3-day period of testing and evaluation. During this period he was continuously under observation and was subjected to a wide range of tests and stresses. In addition to ability tests of a number of kinds—tests of intelligence, mechanical ability, ability to observe and remember details—he was exposed to a number of “situational” tests. These consisted of staged situations, with fairly complete instructions and ground rules, presenting problems that the candidate was to solve, either individually or as a member of a group. The variety of situational tests used in the program was wide. Selected examples are described briefly in the following paragraphs.\*

*The Brook.* Individuals worked in teams composed of five or six men. The group was brought to a stream about 8 feet wide. On the

\* For fuller descriptions and verbatim instructions, the reader is referred to *The Assessment of Men*.<sup>1</sup>

banks were a log, a rock, various boards, ropes, a pulley, and other items. They were instructed somewhat as follows:

In this problem you have to use your imagination. Before you see a raging torrent so deep and so fast that it is quite impossible to rest anything upon the bottom of the stream. The banks are sheer, so it will be impossible for you to work except from the top of them.

You are on a mission in the field, and having come to this brook you are faced with the task of transporting this delicate range-finder, skillfully camouflaged as a log, to the far bank, and of bringing that box of percussion caps, camouflaged as a rock, to this side. In carrying out this assignment, you may make use of any materials you find around here. When the job is done, all of you, as well as any material you have used, are to be back on this side.

The limits within which you are to work are marked by the two white stakes on each bank (the stakes were approximately 15 feet apart), and you are not permitted to jump across the stream.

The behavior of each man in the group was observed, as the group went about solving the task, and each man was rated on such factors as energy and initiative, effective intelligence, social relations, leadership, and physical ability.

*Construction.* A single candidate was presented with the task of building a structure, using materials resembling oversized Tinker Toy. His task was described to him, and then he was told that the test was primarily one of leadership, since the work was to be done by two helpers whom he was to supervise. The "helpers" were called in, and the construction project began. However, the "helpers," who were carefully coached assistants, turned out to be sensitive, stupid, and obstructive, and their behavior ranged from "gold-bricking" to systematically heckling their supervisor. Again, the responses of the examinee to these frustrations were observed, and he was rated for emotional stability and leadership.

*Improvisations.* This test was one of role-playing. Working in pairs, examinees were assigned roles in a dramatic situation and were told to enact the scene as they would handle it in real life. Thus, one situation was set as follows:

A moved to a small city about 3 months ago and opened a business there. He has been doing quite well and one month ago sent in an application for membership in a club in the town. He has heard nothing in response to this application and goes to the home of B, a prominent member of the club, with whom he is pleasantly acquainted. (A is then sent out of the room and B is told that A has received several blackballs. A is then called back into the room.)

As before, aspects of each individual's personality, as they exhibited themselves in the role-playing situations, were rated by observers.

*Stress Interview.* Candidates were instructed to assume that the following situation had occurred.

A night watchman at 9:00 P.M. found you going through some papers in a file marked "SECRET" in a Government office in Washington. You are NOT an employee of the agency occupying the building in which the office is located. You had no identification papers whatsoever with you. The night watchman has brought you here for questioning.

The examinee was given 12 minutes to prepare a cover story to account for his presence in the compromising situation. Then he was subjected to an intensive and grueling interrogation, in which his statements were questioned, inconsistencies brought out, and every attempt made to trip him up and to make him feel foolish. He was rated on the quality of his story and his ability to maintain it and upon his evidence of emotional stability.

Further examples of situational tests might be cited, but these serve to show the essential characteristics of this type of approach to personality appraisal. The attempt is made to develop situations that approach realistic lifelike situations but still permit a reasonable amount of control from person to person. The OSS staff considered desirable characteristics of situations to be that they (1) have a number of alternative solutions, (2) do not require highly specialized abilities, (3) reveal kinds of behavior that cannot be registered by mechanical means, (4) force the candidate to reveal dominant dispositions of his personality, (5) involve interaction with other persons, and (6) require the coordination of numerous components of personality.

#### LEADERLESS GROUP DISCUSSION

One procedure that provides a somewhat simplified version of the situational test, and consequently one that is more widely adaptable for practical use, is the "leaderless group discussion." This approach has been used when a number of individuals are to be appraised for some type of administrative or executive position, such as a school principalship. The candidates are assembled in small groups, a group of about six apparently working best. The group is assigned a topic to discuss or a problem to solve relevant to their background and the position for which they are candidates. They are allowed a substantial block of time—perhaps an hour—to carry on their discussion. During that time they are observed by a team of observers and a

record is kept of the nature and extent of each man's contributions to the work of the group, or summary ratings are made of each group member on those traits and behaviors that can be exhibited in the group situation. A good deal of research has been done<sup>2</sup> on this type of group situation as a personality appraisal device, and the results suggest that the behaviors shown in the limited test situation do have some validity as indicators of life behavior. However, the values and limitations of this type of situational test have still not been completely explored.

#### EVALUATION OF SITUATIONAL TESTS

Situational tests like the leaderless group discussion and those used in the OSS differ from the *May-Hartshorne* character tests in that, though they still deal with behavior in a somewhat disguised situation, they do not yield an actual record or product. Thus, in the *May-Hartshorne* stealing test, it was necessary only to count the coins left in the box to determine the examinee's score. The tests were highly objective as far as the scoring was concerned. Situational tests are not objective. Though an attempt is made to present a relatively standard task situation, the evaluation of each examinee's behavior is through the observations and ratings of the staff of examiners.

The gain from this approach, which offsets the loss in objectivity, is a great increase in the range of behaviors that can be studied. Much that the individual does, especially in his relations with others, leaves no record once the behavior is past. An action showing aggression or resistance to domination, an integrating suggestion that promotes group harmony, assumption of the initiative, or lapsing into passive followership are actions that occur and are gone. We must observe them on the wing if we are to get them at all. This is what the situational test hopes to achieve—to provide the situations that will elicit behavior of this sort and to provide for its immediate observation and rating.

Situational tests appear to be adaptable to eliciting a variety of types of social and emotional behavior that have resisted measurement by any more objective form of test. However, they present a number of problems. A program involving a number of situational tests is costly. The tests are likely to be costly in the facilities and arrangements they require. They are almost certain to be costly in the time of professional personnel to supervise their administration and to evaluate the behavior exhibited in the test situation. The staging of the situations may call for a certain amount of dramatic skill on the part of the examiners, and there is a real problem in

maintaining the uniformity of the situations from individual to individual and from group to group. Another problem is that of preventing leakage of information about the test tasks, so that the task is a novel one to each group as it is tested and is approached by each group with the same background. In view of the practical difficulties involved, it is not surprising that the use of situational tests has been limited to rather elaborate assessment programs, arranged for evaluation of special types of personnel—undercover agents, clinical psychology trainees, or executives and administrators.

The actual value of situational tests and, in fact, of the whole elaborate assessment program remains somewhat of a question. Psychologists who have participated in the programs have been, in many cases, enthusiastic about the procedures. Whether the information that is elicited has real value in predicting important facts about the individual is another matter. In the OSS program, it was possible to obtain only a limited amount of evidence on the extent to which men who had gone through the assessment program turned out well in their job assignments. Ratings from overseas colleagues and evaluations by commanding officers were obtained in a fraction of the cases. Predictions of success did correlate significantly with success on the job. The evaluation that showed the highest correlation was rating for *effective intelligence*. The final rating for effective intelligence based on the complete 3-day program had a somewhat higher correlation with rated success on the job than did scores based on a brief objective test of verbal ability, but the difference was not great.

In another extensive program of situational testing, designed for the selection of clinical psychology trainees\* there was no evidence that the addition of situational tests improved prediction beyond what was possible from the individual's credential file, selected objective tests and a personal autobiography. In summary then, the situational type of test is an interesting additional tool for personality assessment. It seems to provide a direct opportunity to see the individual functioning in lifelike situations and thus to appraise a variety of aspects of leadership, cooperation, and social functioning. However, evidence for the value of the results as improving our prediction of the individual's success on the job is largely lacking. Because its practical value has not been demonstrated and because the techniques are costly in preparations required and in the time of testing personnel, situational testing must be considered a subject for research at the present time, rather than a proven tool for personnel evaluation.



## SYSTEMATIC OBSERVATION

The situational test has introduced us to observation as a technique for studying the typical behavior of the individual. Observation in that instance was of what he did in specified test situations. We turn now to observation in the naturally occurring situations of everyday life. The situations of everyday life are probably less uniform from person to person than the test situations that we stage. Also, they are not loaded to bring forth the behaviors in which we are specially interested. However, the very naturalness of real life events and the fact that we do not have to stage special events just for testing purposes make observation of natural situations appealing to us.

Of course, we observe the people with whom we associate every day of our lives, noticing what they do and reacting to the ways in which they behave. Our impressions of people are continuously being formed and modified by our observations of them. But these observations are casual, unsystematic, and undirected. If we are asked to document with specific instances our judgment that John is a leader in his group or that Henry is undependable, we are usually put to it to provide more than one or two concrete observations of actual behavior to document our general impression. Observations must be organized, directed, and systematic if they are to yield dependable information about an individual.

We should perhaps pause to draw a distinction between the observational procedures that we discuss now and the rating procedures that we considered in Chapter 13. The basic distinction is this: when we are collecting observations, we want the observer to function as nearly as possible as an objective and mechanical recording instrument, whereas when we gather ratings we want the rater to synthesize and integrate the evidence that he has. The one function is purely that of providing an accurate record of the number of social contacts, suggestions, aggressive acts, or whatever the category of behavior may be in which we are interested. The observer serves merely as a somewhat more flexible and versatile camera or recording machine. In rating, by contrast, the human instrument must judge, weigh, and interpret.

Systematic observational procedures have been most fully developed in connection with studies of young children. They seem particularly appropriate in this setting. On the one hand, the young child has not developed the covers and camouflages to conceal him-

self from public view as completely as has his older brother or sister, so there is more to be found out by watching him. On the other hand, he is less able to tell us about himself in words. So it has been in the study of infants and nursery-school children that observational procedures have had their fullest development.

#### STEPS TO IMPROVE OBSERVATIONAL PROCEDURES

Many of the early studies of young children were accounts of the development of a particular child or of two or three children based on observations by a psychologist parent. These provided a general descriptive background for understanding the young child, but they were qualitative and lacking in precision. Careful research with the child or investigations to determine the effect of particular preschool environments or experiences require that we know not merely that he shows negativism and resistance, for example, but also how much or how often. The needs of measurement, as distinct from those of qualitative description, require observational procedures that will permit a statement of quantity, of amount. The procedures should be as objective and reliable as possible, with a minimum of dependence upon the whims and idiosyncrasies of the individual observer. To accomplish this, several precautions are typically undertaken. These are discussed below.

1. *Selecting the Aspect of Behavior to Be Observed.* One problem of the general observer of human behavior is that he does not know what he is looking for. So much is happening in any situation involving one or more active human beings that some part of it must inevitably be missed. We cannot notice everything that happens, and we cannot record everything that we notice. In any program of systematic observation, we must first select certain aspects or categories of behavior to be observed. Thus, in a study of nursery-school pupils, we may be interested in aggressive behavior and may limit ourselves to instances of aggressiveness. In a research project to evaluate a school program, we may be interested in observing evidences of cooperation or of independently initiated activity and may restrict our observation to these.

2. *Defining the Behaviors That Fall within a Category.* If we turn two observers loose without further ado to observe the occurrence of "aggressive acts" or "nervous behavior" in preschool children, we will find that there are many disagreements between them in the observations they make. Our categories must be further specified. They must become more behavioral if we are to get good agreement be-

tween observers. What is an "aggressive act," a "nervous habit"? Do we wish to include name-calling in the first instance? Fidgeting in one's seat in the second? Just as we must analyze "ability to get and interpret data" into specific testable skills of using an index or making inferences from a bar chart, so must we translate "aggressive acts" into hitting, kicking, biting, pushing, grabbing, name-calling, and the like. An advance agreement on what is to be included, based upon prior studies of the domain in question, is a necessary condition of objective and reliable observation.

3. *Training Observers.* Even with a carefully defined set of behaviors to be observed, disagreements arise between observers. Some of these are unavoidable due to fluctuations of attention or variation of scoring on close judgments. Others can, however, be eliminated by training. Practice sessions in which two or more observers make records of the same sample of behavior, compare notes, discuss discrepancies, and reconcile differences provide one means of increasing uniformity. Practice sessions watched and later criticized by an already trained observer represent another. Such procedures make for uniformity of interpretation and standard application of the observation categories.

4. *Quantifying Observations.* If observations of some aspect of the child's behavior, his aggressive acts or his social contacts, for example, are to provide a measurement of the child, some form of quantification is required. The quantification usually takes the form of counting. The count may be of the number of times that a child shows a particular form of behavior during a period of observation. However, in this case one often has difficulty in deciding when one act ends and the next one begins. Johnny slaps Henry and then kicks him. Is this one aggressive act or two? If the actions flow over from one to the other, the decision may not be an easy one.

An expedient that has appeared to work well in a number of cases has been to break the period of observation up into quite short segments. These may be no more than a minute or even half a minute in length. Then the observation that is made is merely the occurrence or non-occurrence of the particular category of behavior during each small segment of time. Thus, we might observe each child for ten 5-minute periods, each on a different day. The 5-minute periods might each be subdivided into ten  $\frac{1}{2}$ -minute periods. For each of the  $\frac{1}{2}$ -minute periods we would observe whether the particular child did or did not exhibit any of a set of defined aggressive behaviors. Each child would then receive a score, with a possible range from 0 to 100, indicating the extent of his overt aggressiveness. Such scores, based

on an adequate number of short samples of observed behavior, have been found to show quite satisfactory reliability. Thus, Olson found<sup>9</sup> the reliability for twenty 5-minute observations of children's nervous habits to be .87 in one case and .82 in another.

5. *Developing Procedures to Facilitate Recording.* An essential for accurate observational data is some procedure for immediate recording of what was observed. The errors and selectivity of memory enter in to bias the reporting of even outstanding and unusual events. In the case of the rather ordinary and highly repetitive events that are observed in watching a child in preschool, for example, an adequate account of what was observed is only possible if the observations are recorded immediately. There is so much to see and one event is so much like others that to rely upon memory to provide an accurate after-the-fact account of a child's behavior is fatal. This is certainly the case in any attempt at complete and systematic recording, though we shall find a place for selective observation and anecdotal recording of significant incidents of behavior some time after they have taken place.

Any program of systematic observation must, therefore, provide some technique for immediate and efficient recording of the events that are observed. There are many possibilities for facilitating recording of behavior observations. One that has been widely used has been to develop a systematic code for the categories of behavior that are of interest. Thus, preliminary observations will have served to define the range of aggressive acts that can be expected from 3- and 4-year-olds. Part of the code might be set up as follows: *h* = "hits," *p* = "pushes," *g* = "grabs materials away from," *n* = "calls a nasty name," and so forth. A record blank can be prepared, divided up to represent the time segments of the observations, and code entries can be made quickly while the child is observed almost without interruption.

If the observer is skilled in standard shorthand, of course, fuller notes of the observation can be taken. These can be transcribed and coded or scored later. In some cases, where a research project has liberal financial backing, more complete photographic or sound-tape recordings of the observations may make possible a permanent record of the behaviors in a relatively complete form. These records can then be analyzed at a later date. Such resources are likely to be the exception, however, and in many cases it will be necessary to plan a simple and efficient code to provide an immediate and permanent record of what was observed. The important objectives here are to do away with dependence on memory, to get a record that will preserve